THE RELATIONSHIP BETWEEN TEST-TAKERS' FIRST LANGUAGE, LISTENING PROFICIENCY AND THEIR PERFORMANCE ON PAIRED SPEAKING TESTS

Suwimol Jaiyote

2016

UNIVERSITY OF BEDFORDSHIRE

THE RELATIONSHIP BETWEEN TEST-TAKERS' FIRST LANGUAGE, LISTENING PROFICIENCY AND THEIR PERFORMANCE ON PAIRED SPEAKING TESTS

by

Suwimol Jaiyote

A thesis submitted to the University of Bedfordshire in partial fulfilment of the requirements for the degree of Doctor of Philosophy

September 2016

THE RELATIONSHIP BETWEEN TEST-TAKERS' FIRST LANGUAGE, LISTENING PROFICIENCY AND THEIR PERFORMANCE ON PAIRED SPEAKING TESTS

SUWIMOL JAIYOTE

ABSTRACT

This thesis presents a study of the relationship between test-takers' first language, listening proficiency and their performance on paired speaking tests. Forty participants from two different L1 backgrounds (20 Urdu and 20 Thai) participated in the study. They took two paired speaking tests: one with a shared L1 partner, and one with a non-shared L1 partner, as well as a listening test and a monologic speaking test to measure their listening ability and individual speaking ability. After each paired speaking test, the participants were also interviewed about their test-taking experience. All speaking tests and interviews were video recorded and transcribed. Raters awarded test-takers analytical speaking test scores (*grammar and vocabulary*, *discourse management*, *pronunciation* and *interactive communication*) and provided comments to justify their scores. Raters also participated in a stimulated recall session. The mixed-methods approach was utilised in analysing and triangulating different data sources. The data analysed in this study included listening and speaking test scores, raters' perceptions of the test-takers' speaking performance gathered from stimulated recalls and test-takers' stimulated recall interviews, as well as the interactional discourse data in the paired speaking formats. The combination of quantitative analysis, Conversation Analysis (CA) and thematic analysis informed the relationship between test-takers' listening proficiency, their L1 and their paired speaking performance.

The results indicated that the greater listening proficiency the test-takers had, the better *vocabulary and grammar* and *discourse management* skills they demonstrated on paired speaking tests. Interestingly, this was observed only when test-takers were paired with a non-shared L1 partner. Similarities and differences in

communication patterns related to interactive listening between the shared L1 pairs and the non-shared L1 pairs were identified. Additionally, other interactional features which are relevant in discussing the differences between the shared L1 pairs and non-shared L1 pairs are also presented in this study.

GLOSSARY

Accentedness is defined as "perceived differences in pronunciation as compared with a local variety" (Munro and Derwing, 2015, p. 14), and is often measured by scalar rating (Munro and Derwing, 2015).

Collaborative task is a type of speaking task in which paired or grouped candidates engage in a discussion without an examiner's intervention. The collaborative task used in this study is part of the *Cambridge First Certificate (FCE)* test, which offers "the opportunity … for the candidates to engage in a discussion and work together towards a negotiated outcome of the task set" (UCLES, 2015, p. 86).

Communication breakdown is a type of miscommunication. It refers to failure to exchange information between a speaker and a listener, and one or both conversants recognise a problem and may attempt to solve it (Gass and Varonis, 1991).

Communication problem consists of two broad types of problematic communication: non-engagement and miscommunication. The former type is defined as talk avoidance or no occurrence of communicative events. The latter type is a disparity between the speaker's intended message and the listener's comprehension (Gass and Varonis, 1991).

Comprehensibility relates to the "perceived degree of difficulty experienced by the listener in understanding speech" (Munro and Derwing, 2015, p. 14), and scalar rating is often used to measure comprehensibility (Munro and Derwing, 2015).

Conversation Analysis (CA) is "the study of recorded, naturally occurring talk-in-interaction" (Hutchby and Wooffitt, 2005, p. 14).

Crosslinguistic influence is the interaction between the previous and following acquired languages (Smith and Kellerman, 1986).

Discourse representation is the overall meaning or the meaning representation of the conversation which is added to the listener's memory. The discourse representation is not fixed and "the listener does not carry in her mind what the speaker actually said but only *her own version of it*. Much depends upon how much the she (the listener) has recalled and how she has interpreted the speaker's words" (Field, 2008, p. 210).

Intelligibility refers to "the extent to which listeners' perceptions match speakers' intentions (actual understanding)" (Munro and Derwing, 2015, p. 14). It is often measured from transcripts, summaries, comprehension questions, identification tasks or true/false verifications (Munro and Derwing, 2015). In spoken language tests, intelligibility represents how easily a spoken language can be understood (Davies et al., 1999), and this study also uses this definition.

Interactive listening is defined as a type of active listening that a supportive listener in a pair (would) offer by providing verbal signs of comprehension or audible support to the speaker (Ducasse, 2010, p. 80).

Meaning representation is "an enriched version of the original piece of information" (Field, 2008, p. 210). It results from the listener using knowledge of the world, the topic, the speaker, etc. in order to comprehend the speaker.

Misunderstanding is a type of miscommunication. It is a mismatch between the speaker's and listener's comprehension of a given utterance. However, no one recognises that a problem has occurred (Gass and Varonis, 1991).

Proposition is defined as "an abstract representation of a single idea" (Field, 2008, p. 209). A listener stores an idea about the speaker's language. A proposition illustrates "the literal meaning of the clause, without regard to its context" (ibid, p. 209).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Name of candidate: Suwimol Jaiyote Signature:

Date:

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr Fumiyo Nakatsuhara, for her continuous support of my PhD study with patience, understanding, and kindness. Her guidance is priceless not only in my study but also my academic career.

I would also like to thank my co-advisors, Dr Lynda Taylor and Dr Chihiro Inoue, and my reviewers, Dr Vivien Berry, Dr Lynnette May and Associate Professor Payung Cedar, for their insightful comments and advice on developing the present study. I am eternally grateful to Naresuan University for being my sponsor and the British Council for partly funding of my study.

Last but not the least, I would like to present my wholehearted gratitude and love to my family, my father, mother and younger sister, and the Phengsiri family for always supporting me mentally throughout my PhD study and my life in general.

CHAPTER 1 Introduction

1.1 General thesis purpose

English is used as a communication tool by numerous people around the world, facilitated by the developments of international transportation, trading and telecommunication. It is used between native speakers (NSs), between native speakers and non-native speakers (NNSs), and between NNSs from different first language (L1) backgrounds. The term "world Englishes" represents the view that English no longer belongs only to English speaking countries but is widely spoken as an international language or "a lingua franca" (Jenkins, 2000). As English is used in international contexts and English language users in different regions tend to speak English with their own accents and/or use different varieties of English, there are some concerns about how well speakers in the combinations of NS–NNS and NNS–NNS comprehend each other's message and achieve their interactional goal. When NSs hear a new accent for the first time, they may take a little time to get used to it and understand it (Buck, 2001). This is more problematic for L2 listeners when they converse with a speaker whose accent is unfamiliar to them (Buck, 2001). It possibly cause problems, disrupt the whole process of comprehension and be followed by communication breakdown. This raises questions as to whether and to what extent effective comprehension and communication rely on their accents and cultural backgrounds, and whether English spoken by the same L1 is more comprehensible to the listener than English spoken by the different L1.

The major motivation of this study is the lack of systematic language testing research on paired interaction discourse that focuses on test-takers' L1 backgrounds. Even though paired speaking formats have been widely used, there are only a handful of studies related to interactions between shared and non-shared L1 test-taker pairs. While there is an increasing number of studies on the impact of test-taker characteristics on test-takers' paired and group speaking scores and features of paired speaking discourse (e.g., Galaczi, 2004, 2014; Lu, 2010;

Nakatsuhara, 2004), research on how non-native speaking test-takers with shared and non-shared L1 interact with the target language in paired speaking tests is rarely found. To fill this gap in the literature, the present study aims to explore how test-takers interact with a shared L1 partner and with a non-shared L1 partner in a paired speaking test. A combination of quantitative and qualitative approaches, called a mixed-methods approach (Creswell and Plano Clark, 2011), is utilised to examine paired speaking test discourse data by focusing on the extent to which the test-takers' listening proficiency correlates with their speaking performance in pairs, and whether their L1 backgrounds (shared and non-shared L1) affect their paired interaction, particularly in relation to their interactive listening.

Methodologically, this thesis explores the nature of co-constructed interactions between NNSs of English, with an emphasis on interactive listening, in paired speaking tests. It investigates the relationship between test-takers' L1, their listening proficiency and their speaking performance in pairs. Differences and similarities in communication patterns related to interactive listening between shared and non-shared L1 test-taker pairs are systematically observed. Three test tasks – listening, monologic and paired speaking – are utilised to assess the test-takers' listening, monologic speaking and interactive speaking proficiencies. Conversation Analysis (CA) is used to reflect the picture of their paired interactions. Additionally, retrospective verbal protocols are employed with the test-takers and the raters in order to gain an understanding of what the test-takers are thinking during their paired performance and to gain further insights into the raters' scoring processes. Data sets obtained from the test-takers' scores in listening and speaking tests, CA, retrospective verbal reports with the test-takers and the raters are triangulated to gain a better understanding of the test-takers' interactions between shared and non-shared L1 pairs.

1.2 Research background

As well as being a "lingua franca" (Jenkins, 2000) among NNSs of English around the world, English is used in daily life in countries such as India and Singapore, and their own standards in English (i.e., World Englishes) have been developed. As

there is a growing number of standard varieties of English, rather than only traditionally recognised standard versions (e.g., standard British and standard American English) as in the past, more research attention has been attracted to the way people in different locations speak because of the impacts of local language and culture, in terms of "its characteristic accents, its syntactic structures, its lexis, its pragmatic features, and the like" (Jenkins, 2006, p. 42).

In 1998, Kachru, who developed the model of the three concentric circles of English, adapted it to the Asian context. The "*inner circle*" is represented by the example of Australia and New Zealand, where people primarily use English as a first language. The "*outer circle*" is exemplified by the cases of Pakistan, India and Singapore, etc., where English functions as an institutional language, and the "*expanding circle*" is represented by examples such as Thailand, China and Japan, where English is primarily utilised as a foreign language. Certain characteristics are shared by these three circles; in particular, all varieties of English are transplanted and constitute the formal and functional distinctiveness of the varieties of English in Asia (Kachru, 1998, p. 93).

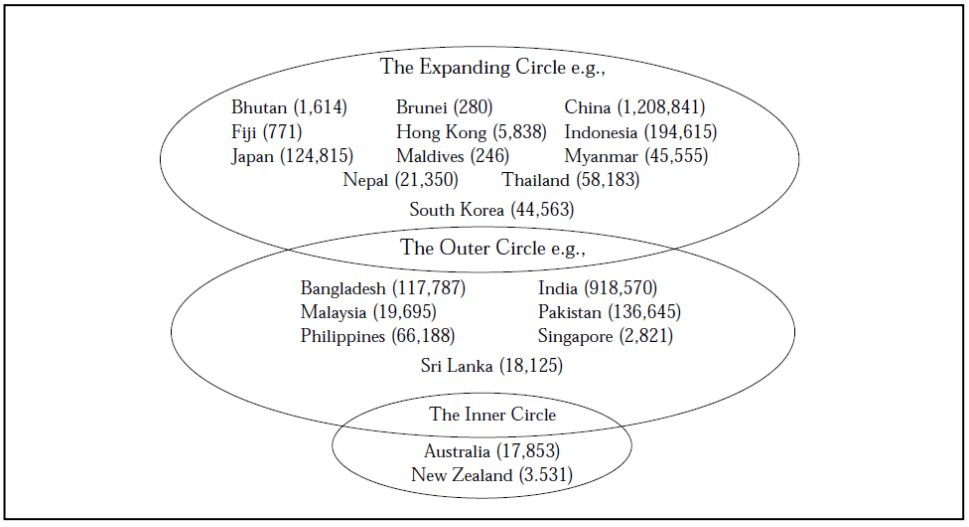


Figure 1.1: Three Concentric Circles of Asian Englishes (Populations in Thousands) (Kachru, 1998, p. 94)

A number of studies provide evidence to support the notion that English spoken by non-shared L1 speakers has an impact on L1 listeners. Non-native

listeners who had the same L1 background as speakers were more able to comprehend the speakers' speech than listeners who had a different L1 background from the speakers' (e.g., Bent and Bradlow, 2003; Harding, 2012; Kachi, 2004). In addition, the listeners who shared an L1 background with the speakers tended to have a better ability to infer what the speakers intended to say based on the linguistic and cultural background knowledge they shared (Kachi, 2004).

In contrast, a listener who does not share an L1 background with a speaker could have difficulty in understanding the speaker's speech, possibly due to cross-linguistic influence, especially on pronunciation, e.g., stress and intonation (e.g., Harding, 2012; Ockey and French, 2014; Stibbard and Lee, 2006). For example, Ockey and French (2014) examined L2 speakers' accentedness, measured by the Strength of Accent Scale, and listeners' comprehension levels, measured by their response to six comprehension questions after listening to a lecture given by speakers with different accents, and reported that the stronger the L2 speaker's accent, the lower the listening comprehension of the L2 listener.

Test-takers' L1 and listening proficiency seem to be important variables which might affect test-takers' interaction in pairs because they need both speaking and listening abilities to achieve their interaction. Therefore, it is necessary to investigate whether non-native test-takers' L1 backgrounds and their L2 listening ability affect their speaking performance in paired speaking test formats. Furthermore, according to the researcher's knowledge, there has been no study systematically investigating the effects of L1 and listening proficiency on both the scores and the discourse of the paired speaking tests. Hence, this study aims to fill this gap in the literature by examining whether and how test-takers' listening proficiency and L1 factors affect their paired speaking performance.

1.3 Rationale for the study

Paired speaking formats, where test-takers are matched to interact with their partners during the assessment, are popularly used as effective tools to assess language learners' interactional ability (e.g., Brooks, 2009; ffrench, 2003; Galaczi, 2004), and it has been demonstrated that a wider range of linguistic functions –

informational, interactional and managing interactions (e.g., ffrench, 2003; Plough et al., 2011) – can be elicited in the format than oral proficiency interview formats (OPIs). An interactional performance elicited from paired formats is also considered as similar to real-life conversation, because the test-takers can control their interaction naturally by changing between speaker and listener roles (e.g., Ducasse, 2010; Ducasse and Brown, 2009; Galaczi, 2014). Other benefits of the paired formats over OPIs include their time-efficiency and cost-effectiveness. As such, the paired formats are widely employed in educational contexts as well as in high- and low-stakes language assessments. They are now used as standard practices in all the Cambridge English for Speakers of Other Languages (ESOL) Main Suite examinations.

The paired format focused on in this study is the *collaborative task* of the speaking component of *Cambridge First*, or the First Certificate in English (FCE) developed by Cambridge English Language Assessment. The FCE was originally introduced in 1939 (UCLES, 2015, p. 3), and the paired format has been obligatory for the FCE since 1996 (ffrench, 2003). The FCE is equivalent to Level B2 of the Common European Framework of Reference (CEFR; Council of Europe, 2001), and the test was taken by over eight million learners all over the world in 2015. The purpose of the FCE Speaking Test is to assess test-takers' ability to communicate effectively in face-to-face situations (UCLES, 2015).

The increasing use of the paired formats in high-stakes examinations such as the Cambridge Main Suite examinations has also called for more research into various issues related to the test format. Assessing speaking performance in paired formats is complex because test-taker performance is related to the candidate's underlying competence and other variables (McNamara, 1996, p. 86), as presented in Figure 1.2.

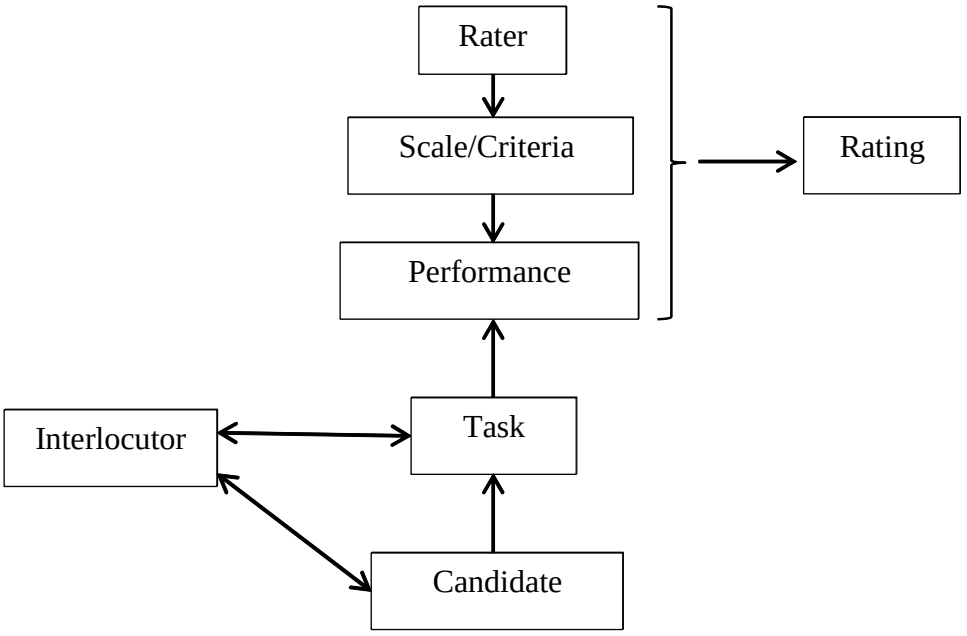


Figure 1.2: 'Proficiency' and its relations to performance (McNamara, 1996, p. 86)

McNamara et al. (2002, p. 228) remind us that "the view of oral test performance as interactive, so central to much current work, means that it is difficult to consider the impact of test-taker characteristics in isolation from those of interlocutors". In paired formats, "an individual's performance is clearly affected by the way the discourse is co-constructed by the person they are interacting with" (Weir, 2005, p. 153), and therefore how to pair test-takers should be carefully considered and appropriately conducted. The test-takers might be treated unfairly in the assessment and this can decrease the validity of the test if possible factors which can affect the test-takers' performance are neglected (Foot, 1999).

To respond to such calls for research, many studies on paired or group formats have thus far identified a number of test-taker characteristics that could affect test performance. They include test-taker personality (e.g., Berry, 1993, 1997, 2007; Ockey, 2009, 2011), language proficiency (e.g., Davis, 2009; Iwashita, 1998; Nakatsuhara, 2006; Norton, 2005), gender and acquaintanceship/familiarity (e.g., Norton, 2005; O'Sullivan, 2002), age (e.g., O'Sullivan, 2008) and L1 (e.g., Jenkins, 1997, 2002; Lu, 2010). While these studies have contributed to our understanding of the role of test-taker characteristics, the impact of test-takers' L1 variable has

still not been extensively researched, especially in relation to the test-takers' listening proficiency. This study aims to fill this gap by investigating the impact of the L1 factor and listening proficiency in paired formats while controlling other variables which might confound the research findings (e.g., age range, gender, overall English proficiency, English speaking proficiency and English listening proficiency) as much as possible. It is hoped that a better understanding will be gained of how the L1 and listening proficiency of test-takers are correlated to their paired speaking performance, whether pairing with shared and non-shared L1 partners presents any similarities or differences in communication pattern, and whether shared L1 test-taker pairs comprehend each other better than non-shared L1 test-taker pairs. The findings of the current study will shed light on paired interaction discourse between a shared and non-shared L1. It is also hoped that the findings will be beneficial for classroom assessment with international students and for high- and low-stakes tests that use paired speaking formats so that the test providers can make an informed decision about how to match test-takers with a shared or non-shared L1 partner, in order to enhance validity and fairness for test-takers.

1.4 Research setting

The current study was conducted at a university in the United Kingdom. The research participants were pre-sessional English language programme students. They are from two different L1 backgrounds (Thai and Urdu) and the numbers of male and female participants were equal. There were two major reasons for selecting the specific participant population for this study. First, they are from different L1 backgrounds; therefore, they possibly reflect the effect of L1 on their interaction in English, on the assumption that L1 backgrounds affect test-takers' speaking performance. Second, the Thai and Urdu languages are from different language families (Thai is a Thai–Kadai language; Urdu is an Indo–Aryan language), and therefore it is less likely that the particulates share L1-related knowledge of each other. As discussed earlier, different L1 test-takers might speak English with accents derived from their own L1, and this can cause difficulty in

understanding each other in non-shared L1 pairs. Consequently, test-takers in non-shared L1 pairs might encounter communication breakdown because of unintelligibility of their partner's speech. While test-takers in non-shared L1 pairs are assumed to face difficulty in their interaction, test-takers in shared L1 pairs are expected to gain some advantages in their interaction from being paired with partners who share an L1 background with them. This study also observes the similarities and differences in communication patterns related to interactive listening in shared and non-shared L1 paired test-takers.

The research questions of the present study are as follows:

RQ1: To what extent is test-takers' performance in paired speaking tests in shared and non-shared L1 pairs affected by their listening proficiency? ;

RQ2: Are there any differences in paired speaking scores when test-takers are paired with shared L1 partners as compared to (when they are paired with) non-shared L1 partners?; and

RQ3: What are the similarities and differences in communication patterns between shared L1 pairs and non-shared L1 pairs?

1.5 Thesis structure

This thesis consists of six chapters. Chapter 1 (Introduction) explains the rationale for the current study, the research setting and the thesis structure.

Chapter 2 (Literature Review) reviews literature relevant to the current research. It contextualises the present study by reviewing theories underpinning the study. Weir's (2005) socio-cognitive framework for validating speaking tests is discussed. Theories relating to L2 speaking performance, interactional competence and speaking proficiency assessment and factors affecting the co-construction of paired/group interactions are reviewed. This study utilises paired speaking test tasks to elicit test-takers' speaking performance. As their speaking performance is co-constructed with their interactive listening proficiency, research on listening comprehension, interactive listening and the impact of L1 on L2 listening comprehension is also reviewed.

Chapter 3 (Research Methodology) explains the research methods used in the

present study. It presents the research questions, the research instruments, data collection processes and the findings of the pilot study. The pilot study aimed to examine, try out and develop the research instruments for the main study. Following the pilot study section, the modified research instruments, research participants and data collection process in the main study are presented.

Chapter 4 (Results of Test Score Analysis) presents the findings gained from the quantitative analysis. The results of the questionnaire which relate to the demographic information about the test-takers, English proficiency based on a standardised examination, and familiarity with English spoken by shared L1 and non-shared L1 speakers are illustrated. The reliability of the listening test and the inter-rater reliability of the speaking test scores are analysed. After that, this chapter presents quantitative results to address the following:

- the relationship between listening and speaking test scores in both monologic and paired formats (RQ1); and
- the impact of test-takers' listening proficiency on performance in paired speaking tests between shared L1 (native language) pairs and non-shared L1 pairs (RQ2).

Chapter 5 (Results of Interactional Data Analysis) presents the results gained from discourse analysis of data together with stimulated recall analysis and discussion. It reports and discusses communication patterns related to interactive listening between shared and non-shared L1 pairs. It interprets and elaborates on the statistical findings by examining the actual interaction by using Conversation Analysis (CA).

Chapter 6 (Discussion and Conclusion) summarises and synthesises the research findings of the present study. It provides the implications of the research findings and the contributions of this study. This chapter also addresses the limitations of the current study and provides suggestions for further research.

CHAPTER 2 Literature Review

This chapter reviews literature relevant to this research, especially focusing on paired speaking test formats and second language (L2) listening comprehension. It consists of seven main sections.

Section 2.1 portrays the theory base for the study.

Section 2.2 focuses on speaking proficiency assessment. Comparative research into oral proficiency interview formats and paired speaking test formats is described.

Section 2.3 presents factors affecting the co-construction of paired/group interactions, i.e., characteristics of test-takers and their interlocutor, and raters' perceptions of co-constructed performance.

Section 2.4 is concerned with listening and the role of interactive listening in speaking test formats.

Section 2.5 explains the impact of L1 on L2 listening comprehension.

Section 2.6 summarises the literature reviewed in this chapter.

2.1 Theory base for the study

Douglas (1998) states that a language test is an elicitation device in the Second Language Acquisition (SLA) which can be designed and interpreted according to a theory of language use. Young (2000, p. 1) argues that

> one of the ways in which language testing interfaces with applied linguistics is in the definition and validation of the constructs that underlie language tests. When language testers and score users interpret scores on a test, they do so by implicit and explicit reference to the construct on which the test is based.

Young suggests that to create a language test, one should be concerned with the definition and validation of constructs underlying the test, based on applied linguistic theory; these issues are presented below.

2.1.1 Construct definition

A *construct* is defined as "a meaningful interpretation of observed behaviour" (Chapelle, 1998, p. 33). It is therefore crucial to discuss how consistently the observed behaviour can be interpreted as an indicator of test-takers' ability. To understand the nature of mediating variables underlying consistent performances in test and non-test behaviours, Messick (1981) identifies three theorist perspectives of construct definition: *trait theorists*, *behaviourists* and *interactionalists*; this has been revisited by Chapelle. Chapelle (1998) defines the perspectives of construct definition as follows.

In defining a construct as a *trait*, a person's consistent performance in a test is related to how a person can use his/her knowledge and underlying process in a test context and apply them in all contexts. Trait theorists view performance as a sign of the underlying characteristics of that person. In defining a construct as a *behaviour*, a person's consistent performance in a test is related to the context of observed behaviour or performance. In this perspective, an individual's performance can be assumed only in the test or a similar context and it will not be generalised to any other contexts. In defining a construct as *interactional*, a person's consistent performance in a test indicates an underlying trait characteristic of that person and the influence of the context in which that performance occurs. Interactionalists view performance as "a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts" (Chapelle, 1998, p. 43).

The interactionalist notion of construct not only covers the idea of the construct from the perspectives of the trait theorist and behaviourist but also mends some defects in both theories. The definition of construct in interactionalist theory is similar to that in the trait theory in terms of the underlying requirement. Knowledge and fundamental processes, including the metacognitive strategies within a specific context, are required in defining construct in both theories. Regarding the issue of test construction, the interactionalist theory is similar to the behaviourist theory. Both theories require the test content to be informed by careful sampling from the context of the target language use. Nevertheless, the

interactionalist is different from the behaviourist in some aspects. The behaviourist seems to believe in surface similarities between the context of the language use in the future, while the interactionalist carefully examines the underlying abilities which are required in the test context.

The current study follows the interactionalist perspective, and it focuses on how and to what extent characteristics of test-takers and their partners (e.g., L1), as part of the performance elicitation contexts, affect their performance in paired speaking tests. The performance of the test-takers is not only affected by their underlying competence but also by the test context, including their partner's L1. Since the definition of construct in the interactionalist perspective is to observe the influence of the test-taker's ability as well as the context of the performance, the understanding of the context is crucial for test validity.

The validation framework selected for this study is Weir's (2005; further elaborated in Taylor, ed. 2011) socio-cognitive framework for speaking tests, which is in line with the interactionist view of the test construct. The framework covers all relevant aspects of the interactionalist perspective by considering an individual's cognitive ability and the social aspects of language use (O'Sullivan and Weir, 2011). The framework is developed to be theoretically informed as well as practically useful, as demonstrated by the wide application of the framework in test development and validation projects by international examination boards, such as Cambridge English Language Assessment and the British Council (e.g., O'Sullivan, 2013; Taylor, ed. 2011). More details of the framework are provided in the following section, which also discusses the notion of validity.

2.1.2 Validity

Validity had traditionally been portrayed in various ways. However, a general consensus in language testing circles seems to have been formed since Messick (1989, p. 13) gave a definition of validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other models of assessment". In short, validity is a concept related to meaning,

interpretations and interferences based on test scores. In Messick's view, the centrality of construct validity and the significance of social dimensions are considered within a unified theory of validity. In this perspective, an overall evaluative judgement of validity evidence is best to display the unified validity of a test.

Following Messick's conceptualisation of validity, Weir (2005) proposed the socio-cognitive framework for test validation, which is now widely used in language testing fields as a comprehensive framework in which validity judgements can be conducted with confidence (e.g., O'Sullivan and Weir, 2011; Taylor, 2011). The framework consists of cognitive processing theory that relates constructs to equivalent operations of language use in real life. In addition, language use in performance tasks is considered as a social phenomenon rather than a totally linguistic phenomenon in this framework. In Weir's socio-cognitive framework for validating speaking tests, he proposes six components which are crucial for creating validity evidence to enhance the test validity. It should be noted that the framework is still evolving; it was modified in 2011 by O'Sullivan and Weir and further modified again in the same year by Taylor (2011). The socio-cognitive framework illustrated in Taylor's research is shown in Figure 2.1.

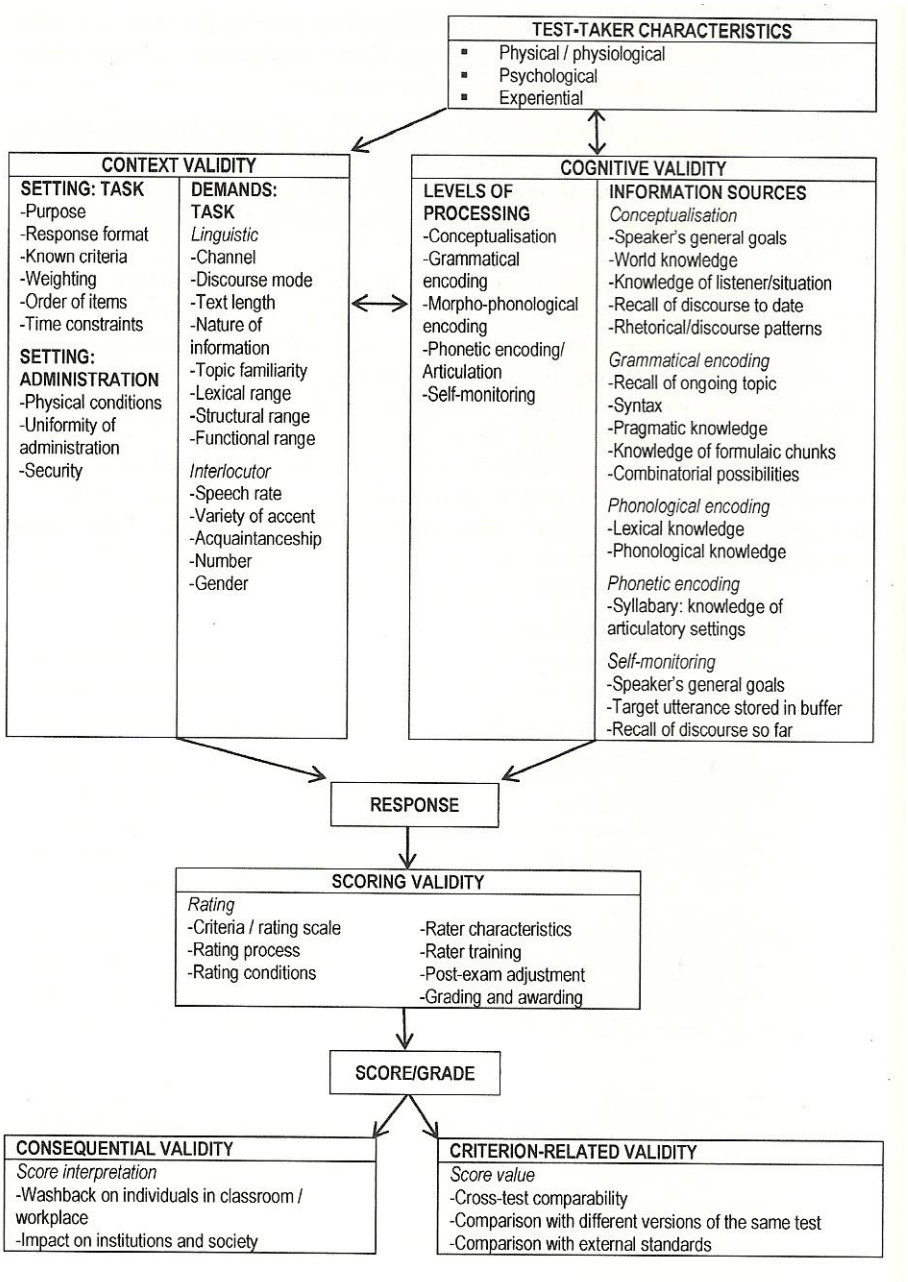


Figure 2.1: Weir's socio-cognitive framework for validating speaking tests as modified in Taylor (2011, p. 28)

The six components in Weir's socio-cognitive framework are *test-taker characteristics*, *context validity*, *cognitive validity*, *scoring validity*, *consequential validity* and *criterion-related validity*. These six components are defined briefly as follows (Taylor, 2011; Weir, 2005):

1 *Test-taker characteristics* is related to how the physical/physiological, psychological and experiential characteristics of candidates are addressed by the test;
2 *Context validity* is related to the extent to which the test tasks are utilised to elicit how the candidates' linguistic proficiency and performance under the testing condition are related to and representative of their real-life performances;
3 *Cognitive validity (*formerly *theory-based validity)* is related to the extent to which cognitive processes which candidates employ to complete the task in the test-setting are similar to what is intended in the theoretical construct and language use in real life;
4 *Scoring validity* is related to the extent to which we can depend on the test scores and the extent to which the test scores are reliable and consistent;
5 *Consequential validity* is related to the impact of the test on teaching and learning and society; and
6 *Criterion-related validity* is related to the extent to which the test scores correlate with an external criterion which is intended to measure the same ability.

These six components are supportive forms of validity evidence. The test validity relates to all facets of validity specified in the framework, and Weir states that interpretation of evidence based on only a single facet cannot be proof of validity of the whole. In addition, it is argued by Weir that providing validity evidence of the test is a responsibility of the test providers in order to ensure that the test is valid to measure candidates' performance as intended and that the cognitive processes required in the test are as consistent as possible with those in non-test contexts.

Weir's socio-cognitive framework is developed to be accessible by researchers and test-practitioners who need to develop and validate tests, since it illustrates various types of validity evidence which test-designers or test-researchers need to gather at each step of the test cycles (Taylor, 2011, p. 25). The arrows in the framework show the directions of the effects of each validity component on related ones. The framework also illustrates the timeline of gathering validity evidence,

which runs from the beginning of the test (the top of the framework) to consequence after the test event (the bottom of the framework). Therefore, researchers and test-practitioners recognise what is necessary at each stage of the test development and how to address test validity in order to make its quality fit with their requirements (O'Sullivan and Weir, 2011). The use of this framework in this study is also motivated by weaknesses of other validation models. For instance, Bachman's model of communicative language ability has been criticised for its difficulty in operationalisation (O'Sullivan and Weir, 2011) and its limitations in addressing interactional and social dimensions (McNamara, 2003).

For these reasons, this study uses the socio-cognitive framework as its theoretical base, and aims to contribute to a better understanding of *test-taker characteristics* (in terms of test-takers' L1 background) and *context validity* (in terms of interlocutor factors and task input) in paired speaking tests, in the hope that the findings will help to clarify the overall validity arguments of paired speaking tests.

To understand the purpose of this research, it is also important to mention two major types of threat to test validity. Messick (1989, p.34) notes that

> tests are imperfect measures of constructs because they either leave out something that should be included according to the construct theory or else include something that should be left out, or both.

The former point refers to *construct under-representation* and the latter refers to *construct-irrelevant variance* (Messick, 1989). The threat of *construct under-representation* is the test failing to cover crucial dimensions of the construct, while the threat of *construct-irrelevant variance* is a test containing variance that is irrelevant to the construct. If a component to assess the construct is not included in the test, the outcome of the test will fail to mirror the construct and it may cause negative washback. On the other hand, if the test includes a component which is not supposed to be measured by the test, scores which test-takers gain may not reflect their real ability. Minimising these two threats is essential to increase positive washback and assure fairness to all test-takers (Messick, 1989). One of the underlying aims of this study is to offer evidence to help us understand how we can conceptualise the L1 backgrounds of test-takers' partners in paired speaking tests

against these threats to test validity. Should they be a part of the test construct or treated as construct-irrelevant variance in paired speaking tests? This question cannot be answered until we have a full understanding of the role the L1 factor plays in paired tests.

2.1.3 Development of theories of L2 speaking performance in relation to interactional competence

Various dimensional models related to communicative competence have been specified in theories of L2 performance, and ability to use language is identified as one outstanding component of communicative language ability in L2 learners (e.g., Bachman, 1990; Bachman and Palmer, 1996; Canale and Swain, 1980). Most recently, interactive competence has been theorised from the preceding theories of competence (Young, 2011). Initially, the models of L2 performance related to spoken language used in face-to-face communication tended to focus merely on an individual's command in communication. For instance, Canale and Swain's (1980) theory of communicative competence focuses on an individual learner's competence in a language in aspects of linguistic, pragmatic discourse, and strategic competence in a social context. Ten years later, Bachman (1990) proposed a model of communicative language ability which consists of language competence and strategic competence. After that, Bachman and Palmer (1996) presented the combination of language knowledge and metacognitive competence in their model of communicative language ability. However, as noted earlier, those models have a weakness as they focus on individual test-takers much more than the interaction of the test-takers (McNamara, 1996). McNamara believes that the language performance of a person tends to be influenced by the test task, by an interlocutor (a person with whom he/she interacts) and by raters who judge the performance.

To respond to such concerns about the theories related to the construct of L2 performance, more attention has recently been paid to research focusing on communication. It is not just a skill possessed by the individual but is also the joint construction of abilities, actions and activities by all participants (Young, 2014, p. 17). The concept of the joint construction by all participants in communication was

first proposed by Kramsch (1986). Kramsch believes that

> successful interaction presupposes not only a shared knowledge of the world, the reference to a common external context of communication, but also the construction of a shared internal context or "sphere of inter-subjectivity" that is built through the collaborative effort of the interactional partners (p. 367).

Later, Jacoby and Ochs (1995) called the joint construction "co-construction", in which abilities, actions and activities related to communicative language are co-constructed by all participants. Interactional competence comprises seven resources which involve knowledge and application of these resources in social contexts that participants use in interaction (Young, 2011):

- Identity resources
    - *Participation framework*: the identities of all participants in an interaction, present or not, official or unofficial, ratified or unratified, and their footing or identities in the interaction
- Linguistic resources
    - *Register*: the features of pronunciation, vocabulary, and grammar that typify a practice
    - *Modes of meaning*: the ways in which participants construct interpersonal, experiential, and textual meanings in a practice
- Interactional resources
    - *Speech acts*: the selection of acts in a practice and their sequential organisation
    - *Turn-taking*: how participants select the next speaker and how participants know when to end one turn and when to begin the next
    - *Repair*: the ways in which participants respond to interactional trouble in a practice
    - *Boundaries*: the opening and closing acts of a practice that serve to distinguish a given practice form adjacent talk (Young, 2000, p. 1).

Advances in research into paired and group tests (see Section 2.2) have found the concept of interactional competence useful in understanding the nature of spoken interaction in these formats (e.g., Galaczi, 2014; Nakatsuhara, 2013). This study also hopes to discuss its results in relation to the assessment of interactional competence, since this study aims to identify the extent to which and in what ways test-takers with different L1 backgrounds bring their L1-related resources to the paired test contexts and deploy them to demonstrate their language-speaking ability.

Keeping this in mind, the next section will detail how paired speaking tests, like the one used in this study, are considered to be capable of assessing speaking ability in interaction. In doing so, various features of the format will be contrasted with those of an interview format.

2.2 Speaking proficiency assessment

In accordance with the shift from the perspective of speaking in a second language (L2) as an information transfer to speaking in interaction, the paired speaking test format was introduced to Second Language Acquisition (SLA) teaching and testing during the 1980s (Ducasse, 2010). Additionally, a more communicative approach in language teaching led to the increasing use of paired work in L2 learning and testing contexts (Taylor and Wigglesworth, 2009) since the paired formats have the potential to tap in a wide range of a learner's interactional ability (Galaczi, 2014). To highlight the advantages of paired speaking tests, a more traditional Oral Proficiency Interview (OPI) format and a paired format are compared in the following section.

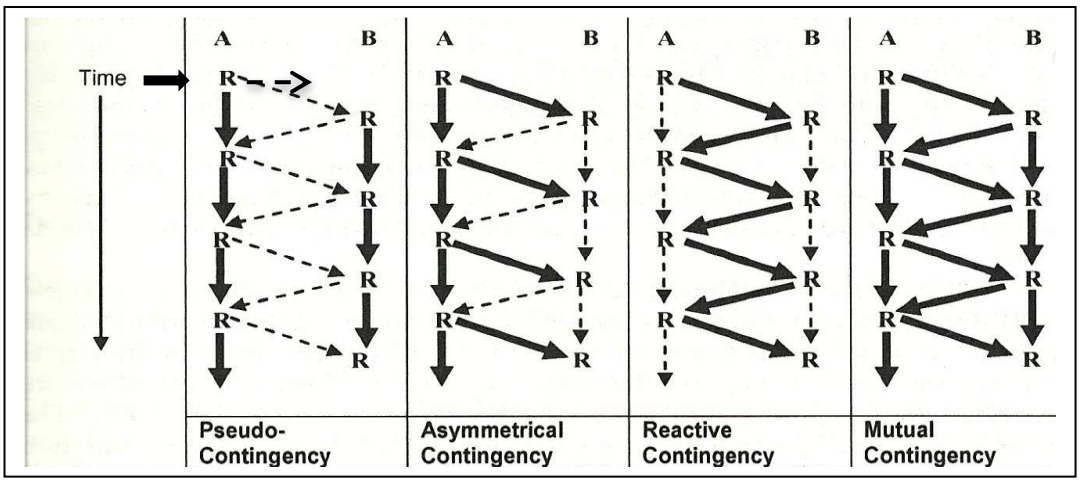
2.2.1 Comparison between oral proficiency interview (OPI) and paired speaking test formats

The paired speaking test format consists of two test-takers engaging in a speaking task and the test-taker's speech performance is co-constructed with their partner. The paired format is widely utilised in assessing test-takers' speaking proficiency because of the limitation of the OPI format in producing rich communicative features. Although it is claimed that "a well-structured oral proficiency interview tests speaking ability in a real-life context – a conversation" (ETS, 1989), numerous researchers, such as van Lier (1989), Johnson (1997), Johnson and Tyler (1998) have criticised the validity of the OPI and the theory related to proficiency which the OPI claims to represent.

Van Lier (1989, p. 494) investigated similarities and differences between interviews and conversations. Two questions were raised to consider whether the OPIs are examples of conversation between people:

- Are OPIs examples of conversational language use?; and
- Is conversational language use the appropriate (or the only, or the best) vehicle to evaluate oral proficiency? (van Lier, 1989, p. 489).

How to characterise the OPIs was discussed using data collected from a) the analysis of various OPIs which he has taken as an interviewee; b) the study of transcripts and tapes of a variety of oral interviews; and c) his past experiences as an interviewer and rater of OPIs of children for diagnostic purposes (pp. 489-490). The model of dyadic interaction of Jones and Gerard (1967) is utilised to distinguish the degree of *interactional contingency* and *goal orientation* in interview settings. Van Lier (1989) defines *interactional contingency* as the involvement of the interactional structure, how participants respond to each other and how various types of sequences create "intersubjectivity". *Goal orientation* means the internal goals of each speaker. Figure 2.2 illustrates the four patterns of dyadic communication of Jones and Gerard.



(A and B represent interactants; R represents a reaction by one interaction to the other; the vertical arrows represent the degree of goal orientation, while the oblique arrows represent the degree of reactiveness and represent primary and minimal (or no) contingency, representatively.)

Figure 2.2: Style of dyadic discourse related to contingency and the orientation of speakers to internal goals (modified from Jones and Gerard, 1967, p. 507 in van Lier, 1989, p. 497)

According to Jones and Gerard's model of dyadic interaction, van Lier (1989) discusses interviews as characterised by *asymmetrical contingency* (one person presents a high degree of goal orientation and the other presents a high degree of

reactiveness); in contrast, conversation normally relates to *reactive contingency* (both parties perform a high degree of reactiveness, but with little goal orientation) or *mutual contingency* (both parties perform a high degree of goal orientation and reactiveness).

Van Lier (1989, p. 498) views that there is asymmetry in the interview test in the exchanges between interviewer and interviewee, since the responsibility for beginning and ending the interaction, ending a topic and introducing a new topic and formulating the talk belongs exclusively to one person. The interviewer has a plan and is responsible for conducting and controlling the interview to follow that plan. In addition, questions are asked by one person (the interviewer) and the other person (the interviewee) is required to answer those questions. According to van Lier's findings, the OPI does not seem to measure speaking ability in the conversation format as it claims to. He also expresses concerns about the misleading use of OPI as a testing instrument to measure the test-takers' ability to carry on a speaking interaction in a real-life context.

In her discourse analysis study, Johnson (2000) analysed data from 35 OPI performances. The OPIs were compared for distribution and allocation of turns in these performances compared with those in ordinary conversation. It was found that turn order, length of turn and turn distribution in OPI interaction are largely fixed and controlled by the testers, who executed them systematically. This is against the salient prototypical characteristic of conversation which OPIs claim to illustrate (Johnson, 2000; van Lier, 1989; Young and Milanovic, 1992). In OPIs, the interviewer is always responsible for selecting the next speaker at every Transition Relevant Place (TRP), while in ordinary conversation responsibility for taking the next turn and managing time within interaction, as well as the rights to select oneself or the other person at TRPs, tends to be equally distributed (Sack et al., 1974). In normal conversation, topics emerge naturally. The topics are negotiated in a process of conversation and are not fixed in advance (Brown and Yule, 1983, p. 89). Additionally, the interviewer's turn consists only of questions and the candidate appears to be allowed to talk only when a response turn is allocated. The findings of the discourse analysis of OPI are in contrast with what the Educational Testing

Service (ETS) claimed as it is found to be similar to a conversation in a real-life context. Johnson (2000) counters that the OPIs do not measure speaking ability in a real-life context-conversation; instead they measure speaking ability in the interview context. Therefore, the validity of the OPI testing instrument is raised as a concern. Measuring test-takers' speaking ability as in a real-life conversation through the OPIs carries the threat of construct under-representation (Messick, 1989) because such tests do not really assess the interactional ability of the test-takers in the conversational context as they intend to do. This concern is confirmed by Johnson and Tyler's (1998) study.

To clarify whether the OPI discourse is similar to natural conversation, Johnson and Tyler examined the discourse in terms of general aspects of everyday conversation, for instance, turn-taking, topic nomination, adjacency pairs and features of conversational involvement. In their study, a Korean female candidate (NNS of English) was interviewed by two American testers (1 male and 1 female). The results showed that the interviewers largely fixed and controlled turn order, turn length and turn distribution of the interaction. This is in contrast with natural conversation, where distribution of turn, order and length of turn are managed and mutually negotiated by people who engage in the conversation. Johnson and Tyler argue that the outstanding features of natural conversation relate to turn-taking and that negotiation of topic seems not to occur in the OPI formats.

Plough et al. (2011) compared two versions (Old test and New test) of the Examination for the Certificate of Proficiency in English (ECPE) Speaking Test. The Old test was in the one-to-one (examiner-examinee) format and the New test is a paired examinee format. Quantitative results indicated statistically significant differences in a range of linguistic functions used by a test-taker on the Old test and the New test. The range of different linguistic functions produced by the test-takers in the New test (34 different functions) was greater than in the Old test (21 different functions). In the Old test, frequently used functions included explanation, general information, opinion, personal information, and hypothesis/speculation. In the New test, the functions of explanation, opinion, and personal information were also frequently used. More importantly, the New test elicited a wider range of functions

than the Old test, for instance, negotiation, presenting, and summarising. In the Old test, there were only two kinds of interactive moves: offer floor and initiate dialogue, in contrast to the New test, where there were five different kinds: take floor, offer floor, request floor, request information, and offer opinion. The findings from Plough et al.'s study provide evidence that the paired speaking format can elicit from test-takers more and different linguistic functions than the one-to-one interview format. That is, the paired speaking format provides the test-takers with the opportunity to demonstrate their ability to engage in discourse, in agreement with an interactionalist perspective of L2 performance, which is what the test is intended to measure.

Similarly, ffrench (2003) analysed the speech functions occurring in the one-to-one interview format and the paired speaking test formats as part of the Certificate of Proficiency in English (CPE) Revision Project by utilising the observation checklist developed by O'Sullivan et al. (2002). The observation checklist consists of three main categories: *informational, interactional* and *managing interaction functions.* The results indicated that 26 out of 30 communicative language functions were found in the paired speaking format, while only 14 functions were observed in the one-to-one format. In addition, it was found that in the one-to-one format, *informational functions* dominated over 80% of all language functions, while all three language functions were distributed quite evenly in the paired format. The approximate percentages of the language function found in the paired formats were as follows: informational functions 55%, interactional functions 30%, and managing interactional functions 15% (ffrench, 2003, p. 414). Details of the distribution of speaking functions in both formats is illustrated in Figures 2.3a and 2.3b.

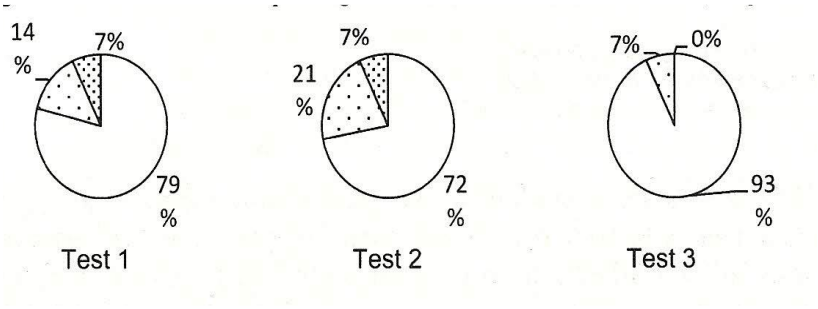


Figure 2.3a: Distribution of speaking functions in the individual format (3 separate instances)

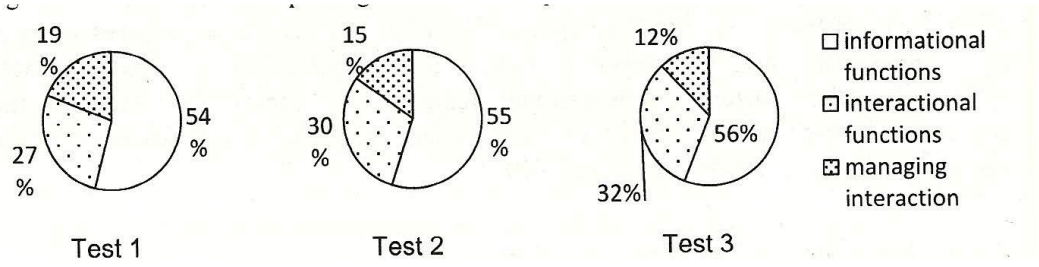


Figure 2.3b: Distribution of speaking functions in the paired format (3 separate instances) (ffrench, 2003, p. 413)

The observation checklist of O'Sullivan et al. (2002) was also utilised to provide the validity evidence of a range of different language functions used in the different formats of the speaking task in the Cambridge ESOL Main Suite examinations (O'Sullivan and Saville, 2000 cited in O'Sullivan, 2008, pp. 175-179). The observation checklist was given to a group of 15 evaluators to apply to a pair of test-takers who participated in four parts of the FCE: *interview*, *individual long turn*, *two-way collaborative task* and *three-way discussion*. The results indicated that informational functions (e.g., personal information, expressing opinions/suggestion, elaborating and justifying opinions) were dominantly observed in the interview task, while only two items from the interactional features were found (i.e., agreeing and responding to a request for clarification). This presents the limitation of the interview format in eliciting from the test-taker interactional functions, which reflects the restricted interactive and principally unequal nature of the interview (O'Sullivan, 2008, p. 175). The collaborative task could elicit all three language functions: informational, interactional and discourse management functions. Therefore, O'Sullivan argued that the collaborative task

could be "truly collaborative in nature" (p. 177).

More recently, Brooks (2009) showed that the paired speaking format could provide students with a better opportunity to demonstrate more proficient English than the one-to-one format. The small-scale study (N=16) compared the test-takers' interaction in two speaking test formats: the individual format (the test-takers interacted with an examiner) and the paired format (the test-takers interacted with another student). The findings exhibited that the test-takers in the paired format tended to get higher scores, displayed a greater range of features of interaction and a more complex interaction, and used more negotiation of meaning between participants than those in the individual format. The qualitative analysis illustrated that the paired format could elicit a wider range of interaction features than the individual format, especially "*prompting elaboration, finishing sentences, referring to partner's ideas,* and *paraphrasing*" (Brooks, 2009, p. 353). In contrast, these four interaction features were infrequent or absent in the performance of students in the interview format.

In addition to the above-mentioned capability of the paired formats in eliciting a wider range of language functions and more real-life conversation than the OPI formats, test-takers also seem to prefer the paired speaking formats (e.g., Együd and Glover, 2001; Iwashita, 1998; May, 2000). Együd and Glover (2001) report that they favour the paired format from their own experience in the Hungarian secondary school context. The positive points of the paired format were as follows:

- students like pairing,
- pairings give students a better opportunity to produce their best,
- pairings help to produce better English than the one-to-one format, and
- pairings support good teaching (Együd and Glover, 2001, p. 70).

May (2000) compared students' reactions when using the OPI and the paired speaking formats. The findings indicated that they preferred the paired speaking test and expressed the view that it gave them the opportunity to exchange ideas, allowing exposure to, and the creation of, new knowledge. Iwashita (1998) suggests that when test-takers perform tasks with a NNS partner, this creates a non-threatening environment compared with a NS interlocutor (interviewer/examiner) and makes them feel more relaxed. Moreover, a positive washback effect of paired

tests in classroom settings has also been reported. Jones (2007, p. 3) notes that due to the introduction of a paired test, when working in pairs in the classroom, students tended to talk more, share their ideas, learn from each other, be more involved, feel more secure and less anxious, use English in a meaningful and realistic way and enjoy communicating in English.

Despite all these desirable attributes of paired speaking formats, research has also suggested that simply pairing students does not always result in desirable, collaborative interaction. In Galaczi's (2004) discourse-based study, 30 paired test-takers' speaking performances in the two-way collaborative task in the FCE speaking test were explored through Conversation Analysis. The aim of the study was to focus on underlying concepts of conversation management, for instance, overall structural organisation, turn-taking, sequencing and topic organisation of the test-takers' interaction in pairs. Global patterns of interaction in the peer test-taker pairs and outstanding interactional discourse features were highlighted in the analysis. The model of dyadic interaction in the ESL classroom setting of Storch (2002), which is based on the dimensions of *mutuality* (the creation of shared meaning from one turn to the next) and *equality* (the distribution of the task among the test-takers) of speech performance, was modified to observe the test-takers' paired speaking performance. Galaczi (2004) identified four patterns of discourse co-construction using CA: *collaborative*, *parallel*, *asymmetric* and *blended interactions*, the first three of which are summarised in Table 2.1.

Table 2.1: Summary of the characteristics of the collaborative, parallel and asymmetric patterns of interaction (Galaczi, 2004, p. 184)

Interactional characteristics	Collaborative interaction	Parallel interaction	Asymmetric interaction	
			Dominant speaker	Passive speaker
Mutuality	High	Low	Low/High	
Equality	High	High	Low	
Topic "life"	Long	Short	Moderate	
Structure of prototypical topic development sequences	A: Topic initiation + topic building ↓ B: Topic extension ↓ A: Topic extension + topic initiation	A: Topic initiation + topic building ↓ B: Minimal acknowledgement + topic initiation ↓ A: Minimal acknowledgement + topic initiation	A: Topic initiation + topic building ↓ B: Minimal acknowledgement ↓ A: Topic extension	

Collaborative interaction is characterised by test-takers who interact with each other with high equality and high mutuality. Both test-takers equally initiate topics and expand their partner's topic, which means that topics are developed over several turns. *Parallel interaction* is characterised by test-takers who work with high equality but low mutuality (a solo vs solo manner). Test-takers in a pair have "equal access to the conversational floor and the development of the task, but are not working together" (Galaczi, 2004, p. 108). They attempt to develop their own topics without extending their partner's topics, resulting in fast topic decay (ibid, p. 185). *Asymmetric interaction* is characterised by "low to medium mutuality and low to medium equality" (ibid, p. 108). This type of interaction is dominated by one test-taker, while the other one takes the passive role and the topics are only developed by the dominant test-taker. Additionally, *blended interaction* is characterised by pairs who show "interactional features characteristic of several interactional patterns, e.g., both collaborative and parallel" (ibid, p. 106).

In Galaczi's data set, the test-taker pairs "oriented either to a collaborative (30%), parallel (30%), or blended (30%) pattern of interaction. The asymmetric dyads were the most problematic from an assessment perspective and they comprised 10% of the data set" (ibid, p. 112). In addition, each of the four

interactional patterns was subdivided into two subcategories by using the conversation dominance (low or high) based on Itakura's (2001) study. The conversational dominance features are characterised as *quantitative dominance* (quantity of talk), *participatory dominance* (interruption) and *sequential dominance* (question). In Galaczi's findings on the distribution of conversational dominance features, it was found that there were some differences between the groups. The test-takers in collaborative pairs mostly utilised "questions as dominance moves, which exercised next speaker selection" (ibid, p. 220). In contrast, the parallel pairs used "interruptions as a conversational dominance move" (ibid, p. 220). The dominance which the collaborative pairs used is termed "*participatory dominance*" and the dominance the parallel pairs used is termed "*sequential dominance*" by Itakura.

Galaczi also discovered a particular relationship between the patterns of paired interactions of the test-takers and their scores for the interactive communication. It was found that the paired test-takers who performed a collaborative interaction pattern normally gained high scores in the interactive communication category, while those who interacted in the parallel interaction pattern (solo vs solo) received low scores. She suggested that the patterns of interaction relate to the level of L2 conversational management ability of the test-takers. That is, the paired test-takers with a collaborative pattern of interaction seemed to have high conversational management ability, and those test-takers with a parallel pattern of interaction tended to have low conversational management ability. Galaczi suggested that the test-takers at a higher level of language development develop "the ability to work with their interlocutor, shift more successfully between the role of listener and speaker, and as such extend the previous turn" (p. 264). On the other hand, the scores of the paired test-takers with an asymmetric pattern were not consistent. The asymmetric pattern was viewed as the pattern which is "simply harder to rate than others because the reasons for the passive and dominant behaviour of the interlocutors may not be straightforward" (ibid, p. 261), so Galaczi requested more rater training for awarding scores to asymmetric pairs.

Based on the studies reviewed above, there is now a general consensus that

the phenomena of turn-taking, topic nomination and topic maintenance in an OPI test are not equivalent to normal conversation, and therefore, "[an] OPI interview cannot be considered a valid example of typical, real-life conversation" (Johnson and Tyler 1998, p. 28). Many researchers support the use of paired speaking formats and confirm that they can elicit test-takers' more interactional abilities (e.g., Brooks, 2009; ffrench, 2003; Galaczi, 2004) and a wider range of language functions (e.g., ffrench, 2003; O'Sullivan and Saville, 2000; Plough et al., 2011), and can show more symmetry between participants than the traditional interview formats can (e.g., Brooks, 2009). In addition, test-takers seem to prefer the paired speaking formats to the OPIs (e.g., Együd and Glover, 2001; Iwashita, 1998), and the positive washback effects of a paired test in classroom settings are reported (Jones, 2007). These are the reasons why the paired speaking or the group speaking formats have become popular for use in classroom and assessment contexts.

As noted above, O'Sullivan (2008) argues that the collaborative task of the FCE Speaking test is clearly collaborative in nature, although Galaczi (2004) adds the caveat that pairing students does not automatically lead to collaborative interaction. The findings of this study might also contribute to such discussions, as the study, using the FCE Speaking test, aims to explore how test-takers' L1 (test-taker characteristics) and their partners' L1 (contextual parameter) could affect resulting interactional performances and test scores. More details on the study design and methodology will be provided in Chapter 3.

2.3 Factors affecting the co-construction of paired/group interactions

While the paired speaking test format has been favoured for various reasons discussed in the previous section, it has also attracted concerns as there seem to be some possibly construct-irrelevant factors affecting the co-construction of the paired interaction. As Luoma (2004, p. 37) notes, the test-taker's speaking performance is co-constructed with their partner's performance, and any of the following may affect the test performance; the test-taker's own and their partner's characteristics, such as personality (e.g., Berry, 1993, 1997, 2007; Ockey, 2009, 2011), language proficiency (e.g., Davis, 2009; Iwashita, 1998; Nakatsuhara, 2006;

Norton, 2005), gender and acquaintanceship/familiarity (e.g., Norton, 2005; O'Sullivan, 2002), age (e.g., O'Sullivan, 2008) and L1 (e.g., Jenkins, 1997, 2002; Lu, 2010). Weir (2005, p. 153) also repeats that "an individual's performance is clearly affected by the way the discourse is co-constructed by the person they are interacting with". In addition, how raters perceive the co-constructed performance in pairs is an additional factor which could affect the test-takers' speaking performance scores (e.g., Ducasse, 2010; Ducasse and Brown, 2009; May, 2007, 2009). As this study focuses on one of the test-taker characteristics, i.e., test-takers' L1 backgrounds and raters' perceptions of shared L1 and non-shared L1 pairs' test performances, this section reviews the literature related to the impact of test-taker characteristics (e.g., personality, proficiency, gender and acquaintanceship/ familiarity, age and L1) and raters' perceptions of co-constructed discourse.

2.3.1 Test-taker characteristics

As portrayed in Section 2.1.2, test-taker characteristics are one of six components in Weir's (2005) socio-cognitive framework for test validation, and partners' characteristics are located under 'interlocutor' as a part of contextual parameters. Although Chapter 1 mentioned a lack of research into the impact of test-takers' L1 background, which is the focus of this study, other test-taker characteristics have been extensively researched. In order to inform the methodology and discussion of this study, it is important to review previous studies on the impact of test-takers' L1 characteristics as well as that of other test-taker characteristics.

2.3.1.1 First language (L1)

As NNSs using English for international communication outnumber native speakers (Crystal, 1997), intelligibility and comprehensibility in using English as an International Language for NNSs has attracted researchers' attention. Research suggests that L1 may affect test-takers' interactional speaking skills and accents (Hahn and Watts, 2011; Isaacs, 2013; Jenkins, 1997, 2002; Lu, 2010; van Engen et al., 2010; Varonis and Gass, 1985a), and Foot (1999, p.37) raises his concern about the factor affecting the validity of paired tests:

> given the likelihood that many candidates will be familiar with the accent of candidates with a different mother tongue, any test which requires candidates to engage in conversation with each other is clearly biased in favour of candidates who share the same mother tongue.

More than 30 years ago, Varonis and Gass (1985a) compared conversational interactions between NS–NS, NS–NNS and NNS–NNS pairs. The NNS–NNS pairs consisted of seven male pairs and seven female pairs. They had either a Spanish or Japanese L1 background and they were also divided by their proficiency levels. The results showed that among three types of pairs (NS–NS, NS–NNS and NNS–NNS), NNS–NNS pairs illustrated the greatest occurrence of non-understanding, while NS–NS pairs showed the least. The NNS–NNS pairs spent more time negotiating meaning than the other types of pairing. When the NNS–NNS pairs were analysed based on their proficiency and L1 background, there were three types: (1) both shared proficiency and L1, (2) either shared proficiency or L1, and (3) neither shared proficiency nor L1. The shared L1 and proficiency pairs showed the lowest occurrence of non-understanding, followed by the pairs in which there was either shared proficiency or L1, and the pairs of neither shared proficiency nor L1 presented the highest. Negotiation was observed mostly in the NS–NNS, followed by NNS–NNS, and then NS–NS pairs. It is suggested that in conversations between NNSs, more negotiated meaning was required because more utterances were uninterpretable. Communication breakdown occurred when participants in a conversation lacked a shared background, linguistic system and specific beliefs, as well as when no attempt was made to negotiate meaning to achieve a communicational goal (Varonis and Gass, 1985a).

Varonis and Gass not only examined the face-to-face interaction in pairs, they also investigated a telephone conversation. Varonis and Gass (1985b) discussed miscommunication in exchanges between NSs and NNSs by focusing on a telephone conversation involving a service encounter between a NNS and a NS. The NNS was assigned to act as a customer who wanted to buy a new television and the NS was an employee of a television repair shop. Their conversation was audio recorded. In order to understand the type of miscommunication which occurred in the conversation, a goal-based model of conversation and a coding

system for interpreting utterances were used. While exchanging messages, participants possibly react to miscommunication in seven ways:

- immediate recognition of problem but no comment;
- immediate recognition of problem and makes comment;
- later recognition of problem but no comment;
- later recognition of problem and makes comment;
- recognition after conversation but no comment;
- recognition after conversation and makes comment; and
- no recognition.

The findings indicate that in this conversation between a NS and a NNS, to keep the conversation going, the NS always changed her belief space to match the perception of the conversational goal of the NNS, while the NNS seemed to use surface conversational devices, with little understanding of the conversation or of the mismatch of conversational goals. Communication breakdown may be inherent in conversations between NSs and NNSs of English since they do not share linguistic and cultural systems for expressing their ideas. Nevertheless, a variety of social and linguistic factors may affect the resolution of a breakdown. The use of negotiation routines is one technique for avoiding conversational breakdown.

In Jenkins' (1997) study, in which she observed paired discourse in the Cambridge CAE, it was revealed that candidates who shared an L1 might try to adjust their pronunciation by using a more L1 accent in order to make it more mutually intelligible.

In order to illustrate empirical evidence that supports the claim of phonological intelligibility in English as an International Language (EIL), Jenkins (2002) analysed three different data sets of NNS–NNS interactions collected in EIL contexts. The first data set drawn from her field data consisted of five communication breakdowns in classroom and social situations. It presented the fact of communication breakdown and phonological sources of miscommunication and non-communication. The second data set was two recorded information exchange tasks. The results showed that phonologically derived breakdowns seemed to be more difficult to solve for language users in NNS–NNS interaction than in NS–NS interaction or in NS–NNS interaction. NNSs with below the bilingual proficiency

level had difficulty using contextual cues to remedy their partner's pronunciation errors. The third data set consisted of one recorded social exchange and three recorded information exchanges. It presented attempts of NNS interlocutors to accommodate (converge) their pronunciation intelligibility for one another. All three were gathered from participants who were identified as being in upper-intermediate (FCE) to low-advanced levels (CAE) of the University of Cambridge Local Examinations Syndicate (UCLES). Jenkins considered these participants, who still had interlanguages, as being at a level of reasonable competence but not fully bilingual in English. She calls EIL interaction interlanguage talk (ILT). Major examples of miscommunication based on pronunciation, "consonant sounds, tonic (or nuclear) stress, vowel length, and non-permissible (according to the rules of English syllable structure) simplification of consonant clusters" (ibid, pp. 87-88) were found. Furthermore, Jenkins also indicates that when pronunciation caused miscommunication in NNS interaction, NNSs of less than bilingual competence tended to focus on the acoustic signal and directly decoded their partner's message from what had been heard rather than using contextual information to clarify the meaning. These deviations from a particular pronunciation, especially in consonant sounds, length of vowel and the placing of tonic stress, result in NNSs' pronunciation being unintelligible to a NNS interlocutor, and the context and co-text did not give much help to clarify meaning (Jenkins, 2002). These students seemed to be operating at the decoding stage of Field's (2008) notion of types of listening behaviour. That is, at this stage the listener perceives just the literal meaning of the speaker's utterance, which is not enough to understand their partner's speech (see Section 2.4 for more details).

More recently, to investigate the effect of sharing an L1 background on speech communication, van Engen et al. (2010) developed the *Wildcat Corpus of Native- and Foreign-Accented English*, which contains scripted and spontaneous speech recordings of 24 native speakers of American English and 52 non-native speakers of English. The participants were paired based on the alignment of their target language and the alignment of their native language backgrounds:

- eight pairs of native English speakers,

- eight pairs of native and non-native English speakers,
- eleven pairs of non-native English speakers with the same L1 background, and
- eleven pairs of non-native English speakers with different L1 backgrounds.

Each pair performed the Diapix task, which is looking at a pair of pictures with ten different spots on each. They were seated back to back and talked to each other to find the different spots on each other's card within 20 minutes. The candidates were measured on their communicative efficiency based on the task completion time and word-type-to-token ratio. It was found that the pairs of native English speakers had the most communicative efficiency and the pairs of non-native English speakers with different L1 backgrounds had the least communicative efficiency. Successful speech communication in a global context depends on the following:

- alignment of the talkers to the target language, and
- alignment of the talkers to each other in terms of native language background (van Engen et al., 2010, p. 530).

Hahn and Watts (2011) explored NNS–NS cases of misunderstanding, which related to NNS pronunciation and how they attempted to repair communication breakdown. Forty-two participants were asked to write *"(un)intelligibility tales"* which they experienced from the native and non-native speakers. The phonological features found in misunderstandings between NNS speakers and NS listeners were vowels, consonants, stress, syllable insertion or deletion, strong or weak syllables, and word boundary confusions. When the misunderstandings occurred, the speakers utilised various strategies to solve the problems: repetition, providing additional information, using non-verbal communication such as pointing, spelling and paraphrasing. The speakers used the repetition strategy the most and the spelling and paraphrasing strategies the least. Hahn and Watts pointed out that the findings of their study helped them to understand the listeners and speakers and how they might utilise their backgrounds, knowledge, and personal and social awareness to interpret a miscommunication that had occurred.

Lu (2010) investigated culture-specific and test-driven impacts on learners' discoursal performance in the FCE Speaking Test. Thirty audio recordings of 62 FCE test-takers from various L1 backgrounds were provided by Cambridge ESOL.

Data were categorised into three groups: Chinese, Italian and Mixed L1. It was found that test-takers' L1 affected the type of the most recurrent discourse features utilised in the same task (e.g., informing, sustaining, prolonging, elaborating and extending). Chinese and Italian test-takers' cultural backgrounds and L1s seemed to significantly affect their types of discourse feature use. In terms of the overall discourse patterns, different types of pairing (with a shared L1 or non-shared L1 partner) did not seem to give the test-takers any advantage or disadvantage.

The other interesting NNS interaction study is Isaacs's (2013) study. Interactional patterns of 84 international engineering graduate students from various L1 backgrounds were examined on a collaborative paired speaking test which was taken from the Business English Certificates. Every test-taker was asked to complete a self-assessment questionnaire before performing the collaborative task and a follow-up questionnaire (with 5-point Likert scales) on the achievement of their interaction, how they understood their partner and how they perceived their partner's understanding of themselves, the effect of their partner's pronunciation on their interaction, and their willingness to work with their partner in the future immediately after finishing their collaborative task. Lastly, each test-taker was interviewed individually about their impressions of the interaction with their partner, their oral communication needs and so on. All test-takers' interactions were transcribed by using adapted CA conventions from the work of Atkinson and Heritage (1984). Galaczi's (2008) interactional classification, i.e., *collaborative (symmetric)*, *asymmetric* (further designated as *dominant* or *passive*) and *parallel* patterns, was used to categorise their interactional patterns. The test-takers' interactional performances were investigated in relation to their target language use in daily life, their speaking proficiency level and how they perceived their performance of the task. The results showed that the parallel and collaborative interactional patterns were the most frequently occurring in the data set, while the asymmetric pattern was the least common. Based on the interactional data analysis and the questionnaire analysis, it was found that dominant speakers in an asymmetric group self-assessed their own L2 speaking and listening proficiency and language use the highest, followed by speakers in collaborative and then

parallel pairs. The self-assessment scores by passive speakers in the asymmetric group were the lowest. The test-takers in the collaborative group were satisfied with their interactional quality the most, experienced the pronunciation and communication difficulties the least, and welcomed working with their partner in the future the most. Perceptions of the interlocutors in the parallel and the asymmetric–passive groups about the interactional outcome and their partner were less positive than those of the collaborative pairs because of their poorer engagement in pairs (e.g., lack of topic extension). The asymmetric–dominant group was the most frustrated with the interaction and rated their ability to understand and to be understood by their interlocutor the most negatively. The passive speakers seemed to be concerned about their own pronunciation because they experienced difficulties in daily speaking communication. In addition, linguistic and cultural factors were observed. For instance, more than 50% of Mandarin speakers were categorised as passive in the asymmetric group, while more than 80% of Indic speakers were classified as dominant. One of the Mandarin passive speakers disclosed that in a meeting in China, only the manager spoke. They could not speak because they had to respect their leader. Therefore, this perspective was a potential influence on their turn-taking behaviour and possibly resulted in the interactional difficulties.

The above studies relating to test-takers' L1 show that it is certain that L1 does affect the test-takers' speaking performance in paired tests. To achieve a communication goal in paired interaction, test-takers need to play both speaker and listener roles. To comprehend what the partner has said is essential for the achievement of the interaction. The test-takers might gain some advantage from being paired with a shared-L1 partner because of being familiar with the accent (Jenkins, 1997, 2002). Therefore, shared L1 pairs possibly understand each other more easily than non-shared L1 pairs (e.g., Hahn and Watts, 2011; van Engen et al., 2010; Varonis and Gass, 1985a). While many studies support the test-takers' interaction benefits in sharing an L1 background, Lu (2010) finds that L1 does not seem to affect the test-takers' discoursal performance as expected. However, none of the above studies looked at the impact of test-takers' L1 backgrounds in

conjunction with their listening proficiency, although Varonis and Gass (1985a) illuminate the importance of test-takers' general proficiency. This study aims to fill the gap by investigating possible interaction between test-takers' L1 and listening proficiency factors in paired tests.

2.3.1.2 Proficiency

The impact of test-taker characteristics other than L1 is now reviewed. The impact of the test-takers' and partners' speaking proficiency level on their speaking performance has been well-researched (e.g., Bennett, 2012; Csépes, 2009; Davis, 2009; Nakatsuhara, 2006; Norton, 2005).

Norton (2005) investigated the effect of the test-takers' proficiency on discourse features in the paired formats in the First Certificate in English (FCE) and the Certificate in Advanced English (CAE) Speaking Tests. The data consisted of video recordings of seven pairs of FCE test-takers, and audio recordings of eight pairs of CAE test-takers from various L1 backgrounds. The findings showed that being paired with a partner with higher linguistic ability was beneficial for lower-level test-takers because they were able to incorporate some of their partner's expressions into their own speech.

Nakatsuhara (2006) examined the impact of proficiency level on conversational styles in the paired speaking tests. Data were gathered from 24 international students from a UK university. They were from various L1 backgrounds. According to their language proficiency level, the participants were divided into two groups: advanced and intermediate. All participants participated in two test sessions in which they were asked to perform two-way collaborative tasks from the CAE test with two partners: one with a partner with the same proficiency level and one with a partner with a different proficiency level. Conversational styles were measured in three patterns: *interactional contingency* (topic ratification), *goal orientation* (topic initiation) and *qualitative dominance* (amount of talk). The results showed that there was no significant difference in overall conversation patterns between conversations with a partner with the same proficiency level and those with a partner with a different proficiency level. However, differential interactional

findings were obtained by CA analysis, indicating that similar conversational patterns were the results of different interactional behaviours, such as the advanced level candidates helping their partner with a lower proficiency level in topic initiation, and intermediate level candidates engaging in more negotiation of meaning. Although these interactional features were identified, Nakatsuhara concluded that the candidates' overall discourse patterns seemed to be that they used similar conversational styles (whether paired with a partner with the same or a different proficiency level).

Csépes's (2009) study also had a similar finding. She investigated the effect of partners' proficiency level on the students' paired speaking performance in the context of a leaving exam at a Hungarian secondary school. Thirty students performed three parallel paired speaking tasks with three partners with different proficiency levels: top, middle and lower proficiency. The core students in Csépes's study were candidates who were at the middle proficiency level and had to interact in pairs with a partner from each of the three proficiency levels. Gender and personal quality variables were not controlled in Csépes's study. All speaking performances were audio recorded. The results showed that the level of proficiency of the core students' partners did not positively or negatively affect the core students' paired speaking test scores.

Davis (2009) examined the influence of interlocutor proficiency on speaking performance within the paired speaking test. A group of 20 first-year students at a Chinese university was investigated. They were divided into two groups: high and low English proficiency levels. The students were asked to discuss the task prompts once with a partner with the same proficiency level and once with a partner with a lower or higher proficiency level. The findings of Davis's study were that interlocutor proficiency had little effect on students' paired speaking scores and the amount of language produced did not always give them higher scores. Only the students with a higher proficiency level gained higher scores because they produced more language; the students with a lower proficiency level received a negative result from producing more language. Furthermore, it was found that students with a lower proficiency level tended to produce more language when they were matched

with a partner with a higher proficiency level, and a collaborative interaction pattern mostly occurred when the students were working with a student with a higher proficiency level. Davis added a caveat that the findings from his study might have been influenced by the small number of participants (20 students) and uncontrolled variables, e.g., gender and familiarity, which were a potential influence on the students' paired speaking performance.

More recently, Bennett (2012) examined the effects of inter-candidate variation in linguistic ability on paired speaking tests in a southern Italian context through pre- and post-test questionnaires and speaking tests. Forty-three candidates responded to the pre- and post-questionnaire and 12 out of the 43 were selected to participate in speaking tests: collaborative tasks and an individual task. All candidates took the individual task and each candidate took two collaborative tasks: one with a partner with the same proficiency level and one with a partner with a different proficiency level. The results indicated that before the tests, 80% of all questionnaire respondents (N=43) and 83% of all test-takers (N=12) who took part in the speaking test section believed that the different ability of their partner affected their speaking performance in the paired speaking tests. However, their beliefs changed after the talking part of the tests. No one believed that the different ability of their partner adversely affected their performance. Moreover, 50% believed that it increased whether working with the partners with lower or higher proficiency levels. There was no significant difference in the test-takers' scores in the collaborative tasks and the individual task. This possibly suggests that the linguistic ability of a test-taker's partner seems not to affect their performance.

Given these study findings, being paired with a partner with a different language proficiency level may not be as much of a problem as some researchers (e.g., Foot, 1999) have been concerned about. However, it is true that non-level specific tests that cover a wide range of candidate proficiency, such as IELTS[1], still do not include paired tasks because of the possible effects caused by proficiency differences in paired candidates (Lazaraton, 2002). Moreover, as reviewed earlier,

[1] IELTS refers to the International English Language Testing System.

Varonis and Gass (1985a) indicate that proficiency could interact with L1 factors when it comes to communicating with different L1 speakers. Therefore, to investigate the impact of test-takers' L1 backgrounds, this research will pair candidates with partners with similar speaking and listening proficiency in English.

2.3.1.3 Personality

The degree of extraversion levels of a test-taker and of their partner in paired speaking tests and of group members in the group speaking tests seems to affect their speaking scores and discourse patterns; however, the results are rather mixed. In the paired speaking tests, extravert test-takers seem to gain higher scores when paired with an extravert partner rather than with an introvert partner (Berry, 1993, 1997), while introvert test-takers provide varied results. In the earlier study of Berry (1993), introvert test-takers seemed not to gain any benefits from being paired with either an introvert or an extravert partner. Nevertheless, in her later study (Berry, 1997), introvert test-takers tended to gain better scores when paired with an extravert partner. In contrast, in the group speaking tests, extravert test-takers seemed to gain higher scores when working with introvert group members than with extravert group members (Berry, 2007; Ockey, 2009). The impact of the degree of extraversion on introvert test-takers' speaking performance provided varied results. They possibly received higher speaking test scores when performing with extravert group members (Berry, 2007) or may not have gained any effect by the way they were grouped with other group members (Ockey, 2009).

Reviewing studies related to personality reveals that the effect of personality on test-takers' speaking performance appears complex. It can mean that the personality factor alone is not enough to predict the occurrence of systematic differences in test-takers' speaking performance, as suggested by Berry (1997).

2.3.1.4 Acquaintanceship/familiarity and gender

Foot (1999) expresses his concerns about the influence of matching with a known and an unknown partner. He claims that talking to a friend or someone a test-taker knows might be less stressful than talking to a stranger and so a candidate might

gain higher scores in the former case. To clarify this concern, the effect of acquaintanceship on paired speaking performance has been investigated, often together with another test-taker variable, the gender of test-takers (Norton, 2005; O'Sullivan, 2002).

It can be seen from the findings of O'Sullivan (2002) and Norton (2005) that gender and acquaintanceship seem to affect test-takers' speaking performance and test scores. Test-takers, especially female ones, seem to be sensitive to the gender of their partner. Being paired with a male partner might change their interactional performance, for example, using more back-channelling and allowing their male partner to speak first (Norton, 2005). Female test-takers tend to gain higher scores when paired with a friend rather than a stranger (Norton, 2005). However, the gender variable seems not to affect the test-takers' speaking score when they are paired with a stranger (O'Sullivan, 2002). Therefore, whenever utilising the paired speaking format, especially in the assessment context, it seems advisable to take partners' gender and acquaintanceship variables into account.

2.3.1.5 Age

It has been suggested that in an interaction the relative age of interlocutors might affect their performance (e.g., Bachman, 1990; Weir, 1993). However, there are only a handful of studies related to the effect of age on test-takers' speaking performance (O'Sullivan, 1995, 2008) and the results are inconsistent. While O'Sullivan's first study (1995) found this effect only in Arab test-takers, who gained higher scores when paired with an older partner, the findings of his preliminary study in 2008 found that the age variable seemed not to affect the test-takers' speaking performance. Furthermore, the main study discovered the impact of age on the test-taker's speaking performance. It is not clearly confirmed whether the age variable does or does not affect the speaking performance of the test-takers in the paired formats.

To sum up, the characteristics of test-takers and their interlocutor (e.g., personality, proficiency, gender and acquaintanceship, age and L1) could affect their speaking performance and/or speaking scores in paired or group speaking

formats, although the degree and the direction of effects are not confirmed. It is not clear either which variable has the most impact on the test-takers' performance. Therefore, these variables should be carefully considered when conducting research using paired and group speaking formats. As the current study aims to investigate the effect of a different L1 on NNS–NNS interaction in pairs, other test-taker variables which might affect their performance are controlled, for example, age range, gender, English language proficiency (especially speaking proficiency) and acquaintanceship, with the hope of preventing confounding the findings and offering a specific contribution to the contextual parameters and test-taker characteristics of Weir's (2005) socio-cognitive framework described in Section 2.1.2.

2.3.2 Raters' perceptions

Not only the test-taker characteristics but also the raters' perceptions of test-takers' performance affect test-takers' speaking scores (Ducasse, 2010; Ducasse and Brown, 2009; May, 2007, 2009). This relates to the scoring validity component of Weir's socio-cognitive validity framework.

May (2007, 2009) examined how trained raters would rate the features of a paired candidate in an English for Academic Purposes (EAP) speaking test. Twelve pairs, six pairs at the same level and six pairs at a different level, had their performance video recorded in paired speaking tests. The raters were trained to award interactional effectiveness scores in terms of three aspects: how much the interlocutor's message was understood, the ability to respond and the communicative strategy use. The authentic evidence of understanding a partner was assessed through what the test-takers' response to a partner was, as described at Band 5 of the interactional effectiveness criterion, which is to completely understand the partner's message and respond accordingly. Understanding in the context of an EAP discussion task based on the interactional effectiveness criterion consisted of two aspects of comprehension: (1) the surface decoding of the message and (2) concepts, arguments and the link between ideas. Responding to a partner could be evidence of comprehension and/or engagement with a partner's ideas,

which normally occurs in genuine discussion. May found that the test-takers were awarded low scores for interactional effectiveness by the raters when they showed a limited ability to respond or responded incoherently, minimally responded to their partner or irrelevantly responded, could not respond to their partner's question or point, or did not listen to a partner, as evidenced by interruptions. The shared L1 test-takers could understand each other's message, but it was incomprehensible to the raters. This could indicate that test-takers' L1 affected the way in which they pronounced their words in L2 in terms of accent and the way in which they understood accented L2 produced by their NNS partners. As such, sharing an L1 background could facilitate their comprehension in L2.

Although the raters were trained to award the interactional effectiveness score, it was found that they employed more detailed features which were not indicated in the criterion, for example, body language and assertiveness through communication, and these two aspects were related to culture and L1 usage of test-takers. Therefore, May proposed combining these two aspects when awarding an interaction effectiveness score for test-takers in the paired speaking test format. Because of the difficulty in assessing test-takers' interactional effectiveness, sharing the score for interactional effectiveness with paired test-takers' performance and rating scores in other categories for the individual test-taker were recommended. However, the issue of rating the joint scoring for interactive effectiveness does not seem to be straightforward. For instance, Nakatsuhara's (2009, 2013) study on conversational styles in group oral tests indicated that shared scoring would not always guarantee fairness to candidates, because there were some cases in which test-takers did not achieve collaborative interaction even when some test-takers tried very hard to scaffold their quiet group members' cooperation. This supports the current practice of the Cambridge Main Suite examinations, in which individual scores are awarded to the interactive communication category. To ensure fairness to all test-takers, this study will follow Cambridge's current practice, and award individual scores to all rating categories.

Raters' perception of test-takers' interactional performance was also examined by Ducasse and Brown (2009). How raters evaluated paired test discourse in a

discussion task carried out by learners who were studying Spanish as a foreign language in a university was investigated. The full study was also published in the work of Ducasse (2010). The data consisted of video recordings of 17 pairs of test-takers. Test-takers were asked to watch video recordings of their performance to stimulate verbal recall of their thoughts during the paired speaking test performance. The 12 raters were selected to observe and record comments on each video recordings individually. Each rater commented on three assigned paired test-takers, and each test-taker pair was observed by at least two different raters. Think-aloud protocols were used to gain insight into raters' views on interactional features when they were not guided as to what interactional features they should consider. The results showed that raters awarded interactive speaking scores to the test-takers based on three main categories: *non-verbal interpersonal communication*, *interactive listening* and *interactional management*.

The first category, *non-verbal interpersonal communication,* consisted of two subcategories: *gaze* and *body language*. Positive scores were awarded to test-takers who looked at each other during the interaction, and negative scores were awarded to test-takers who demonstrated lack of gaze. Ducasse (2010) and Ducasse and Brown (2009) notes that even body language might be viewed as a positive factor in achieving communication; it might be viewed negatively when it is used to reduce difficulties in delivery meaning because raters might think the test-takers lack verbal resources. Furthermore, raters viewed non-verbal language as related to some extent to the test-takers' cultural specifics. The second category, *interactive listening,* illustrated test-takers' attention or engagement while listening to their partner during the interaction. It was divided into two subcategories: *comprehension* and *supportive listening*. More details related to this category are presented in Section 2.4.1. The third category, *interactional management*, was considered through how the test-takers managed the topics and turns. It consisted of two subcategories: *horizontal* and *vertical management*. *Horizontal management* could be viewed through how test-takers managed between adjacent turns that made the conversation flow and components which related to speaker change, for instance, response speed, length of turn or turn domination. The second category of

interactional management is *vertical management* which is viewed through how the test-takers managed to connect topics to complete the conversation. It illustrates flexibility of the test-takers to allow switching between topics. Raters viewed both turn change and topic cohesion as indicators of successful interaction. Since it is very complicated to judge the test-takers' interactional competence, Ducasse and Brown proposed that language-testers should seriously consider the facets of a construct for a paired interaction to cover non-verbal communication and listening, as well as turn and topic management as the major characteristics.

From reviewing the studies related to raters' perceptions, there is strong evidence to confirm that how the raters perceive test-takers' performances does affect their scores, especially for interactional effectiveness. It seems to be crucial to introduce more aspects, which are not included in the criterion, into the interactional effectiveness score, for instance, body language (Ducasse, 2010; Ducasse and Brown, 2009; May, 2007, 2009), assertiveness (May, 2007, 2009) and evidence of comprehension (May, 2007, 2009) or interactive listening (Ducasse, 2010; Ducasse and Brown, 2009).

To summarise the literature review thus far, although there are various concerns related to test-taker characteristics in employing the paired speaking formats (Foot, 1999), as Együd and Glover (2001, p. 76) state, "[T]he paired formats offered students and teachers opportunities for development, and an escape route from the prison of dire one-to-one situations". Compared to the OPI formats, using the paired speaking test formats in language assessment and teaching can provide many more advantages than disadvantages, for example, eliciting more symmetrical contributions to the interaction and richer and more varied language functions (e.g., Brooks, 2009; ffrench, 2003; O'Sullivan and Saville, 2000; Plough et al., 2011), fostering a positive washback effect to the classroom and positive feedback from test-takers (e.g., Együd and Glover, 2001; Jones, 2007, May, 2000). However, it is true that a number of factors are involved in paired speaking tests, and they should be employed with care by trying to control possible factors which might impact on the test-takers' speaking performance or scores as far as possible. The present study therefore not only concerns the impact of the characteristics of test-takers and their

interlocutor on their interactive performance in pairs (as stated earlier in Section 2.3) but also looks at how raters award scores for shared and non-shared L1 pairs. This study therefore aims to investigate the impact of test-takers' L1 background comprehensively, touching upon three components of Weir's socio-cognitive framework*: test-taker characteristics*, the interlocutor factor as one of the *contextual parameters* and rater perception as a part of the *scoring validity* parameters.

As noted earlier, to what extent test-takers understand their partner when in a pair is crucial for successful interaction in paired speaking tests. This means that test-takers' listening ability, as well as their speaking ability, plays an important role in the test format. In the next section, some of the literature on listening comprehension will be reviewed, with a special focus on the listening ability required during interaction.

2.4 Listening comprehension

Listening is often believed to be the least explicit language skill, resulting in it being the most difficult skill to learn (Field, 2008; Vandergrift, 2004). Listening comprehension is an active process (Buck, 2001; Vandergrift, 2004) of constructing meaning and applying knowledge to the hearing of sound (Buck, 2001). When listening, listeners utilise their background knowledge of the world to build up expectations, which are later used to assist them in understanding what they hear (Buck, 2001). Nevertheless, L2 listeners with a below-bilingual proficiency level seem to fail to utilise background knowledge or contextual information to help them understand their interlocutor's speech (Jenkins, 2002).

While listening, a listener uses two types of listening behaviour: *decoding* and *meaning-building* (Field, 2008, p. 85). The former listening behaviour is the way in which a listener deals with what is heard. During this process, the listener perceives just the literal meaning of the speaker's words, which it is not enough to understand what the speaker has said; therefore, the meaning-building process is required. In the meaning-building process, the listener interprets the meaning of the speaker's speech by utilising external information, for instance, knowledge of the world. This

process consists of three levels: *proposition*, *meaning representation* and *discourse representation* (see Figure 2.4; see also Glossary).

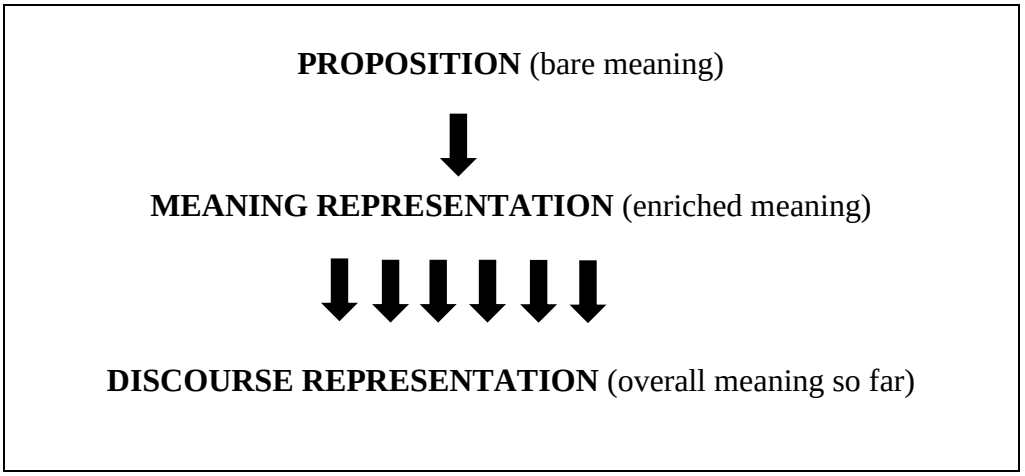


Figure 2.4: Three levels of meaning (Field, 2008, p. 210)

High automaticity of the cognitive process, e.g., decoding messages as well as composing spoken contributions, is essential to being an efficient listener in conversation (Field, 2011).

2.4.1 Interactive listening

Listening is emphasised as one of the abilities which contribute to the achievement of an effective interactive spoken language performance (ffrench, 2003; Galaczi, 2004). Shifting between the roles of a listener and a speaker is key for successful interaction (Galaczi, 2004). Responding to the speaker is the listener's responsibility (Buck, 2001), and how listeners respond to the speaker has been used to observe the listener's interactive listening ability (Ducasse, 2010; Ducasse and Brown: 2009). Providing back-channelling is another method used by the listener to present his/her interactive listening. Back-channelling is a sign which a listener gives to a speaker in order to demonstrate that he/she is listening and understanding what is being said, and that he/she is paying attention (Buck, 2001, p. 13).

Interactive listening has been investigated by Ducasse (2010) and Ducasse and Brown (2009) through paired speaking formats. In the paired speaking formats, test-takers have two roles to play, that of a listener and that of a speaker (Ducasse and

Brown, 2009). They need to listen to their partners in order to attend to, or engage in, the conversation. Ducasse and Brown claim listening is a significant factor for successful interaction, and they call this type of listening *interactive listening*.

Interactive listening is divided into two subcategories: *comprehension* and *supportive listening* (Ducasse and Brown, 2009). The former subcategory refers to the test-takers' understanding of their partner's message and their showing comprehension through verbal support, which are a listener's means of showing engagement, encouraging a speaker to continue speaking or demonstrating comprehension (Ducasse, 2010). Whether the listener test-takers understood their partner was observed through these outstanding behaviours: *filling a silence by providing vocabulary* and *asking for clarification or comprehension*. Assisting their partner by providing a word when he/she was searching for one was an indicator of their attention and that they understood sufficiently to expect a missing word, and this performance enabled the interaction to continue. Those test-takers who were listening but who did not engage in the conversation or support the speaker by either signalling the speaker to continue or by taking the speaking floor and offering to break the silence were considered to be un-interactive or to be engaging in "unsuccessful listening" (ibid, p. 77).

The second subcategory, *supportive listening*, is verbal signals of understanding or audible support (i.e., back-channelling) and sometimes non-verbal communication (e.g., gesturing) used by a listener during a conversation to encourage the partner to continue speaking and maintain the floor. The use of back-channelling and gesturing were signs of interactive listening and negotiating comprehension between the test-takers. Nevertheless, this type of listening does not always necessarily illustrate that a test-taker understands their partner's speech. Sometimes, the test-takers would pretend that they were listening until they understood, or even if they were not listening, they would pretend to be listening but instead would be thinking of how to deliver their idea.

Ducasse (2010) argues that interactive listening during speaking, non-verbal interpersonal communication and demonstration of speaking engagement through interactional management are important for successful interaction. In the paired

speaking formats, interactive listening skills, such as showing evidence of comprehension, were key interactional factors for successful interaction (Ducasse, 2010; Ducasse and Brown, 2009).

Achieving interactive communication is effectively shifting between the roles of a speaker and a listener, and it is related to how proficient test-takers are in interactive listening. Galaczi (2004) investigated the interaction of test-takers in a paired speaking test in the case of the FCE (see Section 2.2.1 for more details). The results showed that test-takers with a higher proficiency level could shift between the role of the speaker and the listener more successfully than those with a lower proficiency one. Those who gained a high score in interactive management were in collaborative pairs, in which the paired members collaborated in developing their own topics and supporting the development of their partner's topic. Galaczi argues that the topic extension move is important in the L2 assessment context since it is evidence of how a test-taker has sufficient language ability to comprehend his/her interlocutor' speech production and to respond by extending the topic. It displays understanding of the previous talk because the test-takers not only have to produce the talk well but must also understand their interlocutor well. That is, they have to take both speaker and listener roles and shift between roles effectively.

The significant role of listening in speaking tests has also been suggested in other interactive speaking test formats, such as group oral and interview tests. In group oral tests, communication problems can occur due to the test-taker's limited listening proficiency, which sometimes leads to the use of meaning negotiation (Nakatsuhara, 2011). The relationship between the test-takers' listening proficiency and their performance in the interview test was investigated by Nakatsuhara (2011). Thirty-six pre-sessional students at a UK university participated in her study. They took Part 2 (Individual long turn) and Part 3 (Discussion) of the IELTS speaking test, a listening test and a short semi-structured interview. It was found that the Fluency and Coherence scores in Part 2 were significantly higher than in Part 3. The cause of the differences in the Fluency and Coherence scores between the two parts seemed to be associated with the additional listening demands in Part 3 of the test. The listening demands resulting in filled and unfilled pauses before responding to

the examiner seemed to have dropped scores for this criterion. According to the students' speaking performance in Part 3, five patterns of communication breakdown related to test-takers' difficulty in understanding questions posed by the examiner were identified by employing the CA methodology. The communication problems, which related to the test-takers' listening proficiency and the ways those problems were dealt with, were classified as follows:

Type a) asking a question and then responding relevantly,
Type b) asking a question and then responding irrelevantly,
Type c.1) understanding a question and responding very irrelevantly,
Type c.2) misunderstanding a question and giving somewhat related but mostly irrelevant response,
Type d) echoing uncomprehended parts, and
Type e) saying "no" to an uncomprehended question (ibid, p. 151).

The results of Nakatsuhara's study showed that the correlations between the students' listening scores and their speaking scores were higher in Part 3 than in Part 2. She argued that Part 3 of the IELTS examination could tap into the students' listening-into-speaking abilities to some extent. In particular, those who had IELTS speaking scores at Band 5.0 or below could not always discuss with an examiner effectively because of their limited listening comprehension proficiency, not just because of their speaking proficiency. However, it seemed that those who were above Band 5.0 rarely had listening comprehension problems with the examiner's questions.

It can be seen from reviewing the related studies above that both speaking and listening proficiencies are necessary for the success of the interaction in interactive speaking formats. To some extent this suggests that these formats are not only assessing speaking skills, but also listening skills, and this point is supported by Galaczi (2014). In Galaczi's (2014) study of the paired speaking interaction co-constructed by learners with different proficiency levels, 41 video-recorded test performances of an interaction task at CEFR levels B1 to C2 were analysed. The findings showed that the candidates with different proficiency levels exhibited three of the most salient features differently. Those features were:

- topic development organisation and specifically:
  - degree of topic development
  - topic extensions of "own" vs "other" topics
- listener support moves:
  - back-channelling (e.g., "*yes*", "*mm*")
  - confirmation of comprehension (e.g., "*absolutely*", "*exactly*")
- turn-taking management:
  - in a no-gap-no-overlap manner
  - following an overlap/latch
  - following a gap/pause.

In terms of *topic development*, the test-takers with a low proficiency level illustrated low mutuality and produced short topical sequences. They predominantly initiated their own topics and rarely contributed to develop their partner's topic initiation. This is in contrast with the test-takers with a high proficiency level, who engaged more with their partner's topics, extended topics over turns and initiated their own topic. As discussed above, topic extension moves are important signs of whether test-takers have sufficient language ability to comprehend their interlocutor's message and to extend the topic further (ibid, p. 562). The difference between proficiency levels is presented in terms of the mean percentage of types of topic development moves in Figure 2.5.

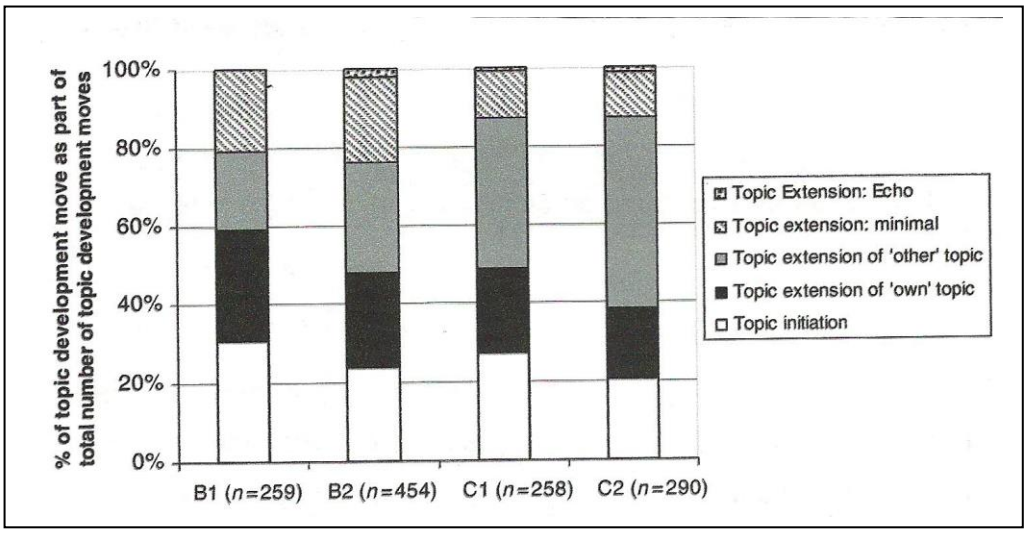


Figure 2.5: Distribution of topic development moves
Note: The *n* counts indicate the total number of topic development moves observed at that proficiency level (Galaczi, 2014, p. 570).

Regarding *listener support*, the test-takers with a low proficiency level provided their partner with limited listener support. They seemed rarely to use listener support and had difficulty in performing both the speaker and the listener roles actively because the paired interaction had high cognitive demands in relation to decoding their partner's speech and producing their own messages (ibid, p. 562). The distribution of listener support moves is presented in Figure 2.6. As illustrated in the figure, the test-takers with the lowest proficiency level provided little listener support and back-channelling (they showed listener involvement, but this did not necessarily mean comprehending their partner) was mostly used, while the test-takers with the highest proficiency level played a role as supportive listeners by providing both back-channelling and confirmations of comprehension. The test-takers with the highest proficiency level used the confirmations of comprehension, which signalled listener involvement, the most. As the confirmations of comprehension increased in use, the back-channelling decreased.

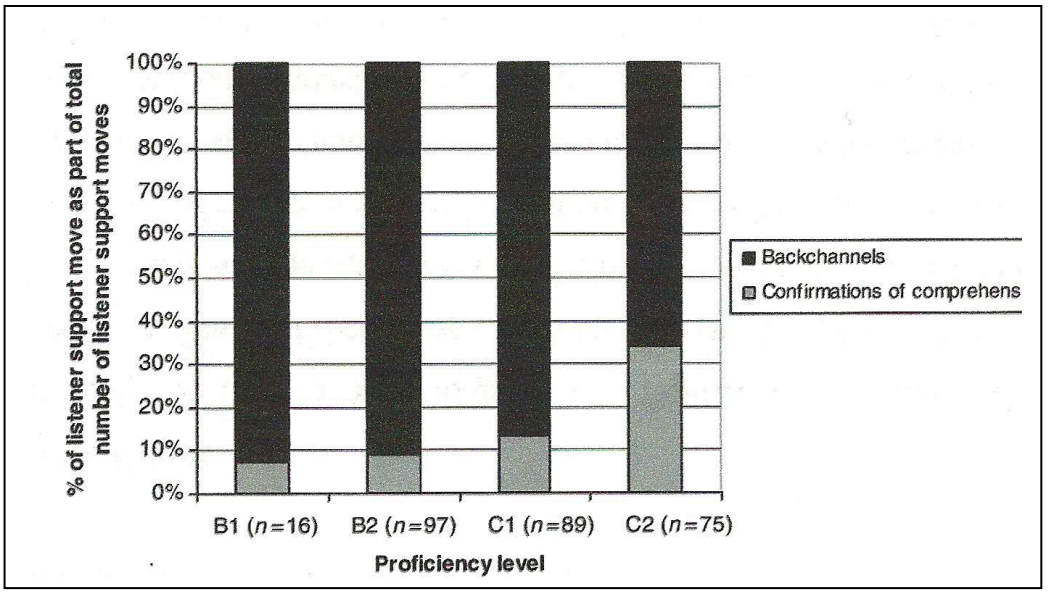


Figure 2.6: Distribution of listener support moves
Note: The *n* counts indicate the total number of listener support moves observed at that proficiency level (Galaczi, 2014, p. 572).

In terms of *turn-taking management*, the higher the proficiency level of the test-takers was, the faster they could start a turn after a latch or overlap (see Figure

2.7).

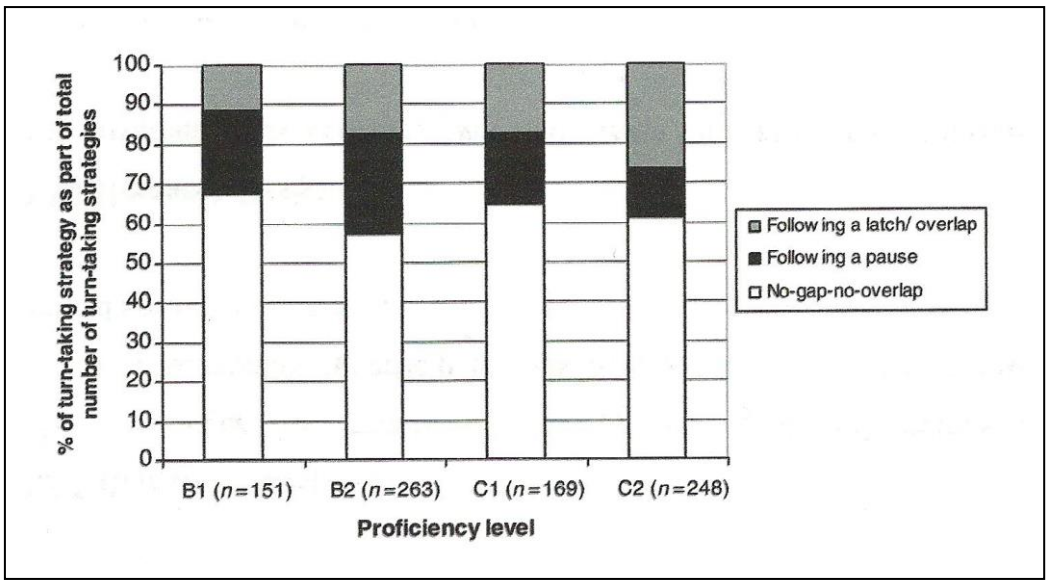


Figure 2.7: Distribution of turn-taking strategies
Note: The *n* counts indicate the total number of turns observed at that proficiency level (Galaczi, 2014, p. 572).

The findings of Galaczi's study suggest that the listener support strategies and turn-taking management should be incorporated into the assessment scales of interactional competence.

The studies reviewed thus far suggest that there is enough evidence to confirm that interactive listening is an important feature of successful speaking interaction (Ducasse, 2010; Ducasse and Brown, 2009; Galaczi, 2004, 2014; Nakatsuhara, 2011). However, there are still only a handful of studies related to this issue. Moreover, the issue has not been investigated in the paired formats together with the impact of the test-takers' L1 on their paired speaking interaction in the shared and non-shared L1 pairs. Therefore, this study will fill this gap in the literature by taking into account the impact of test-takers' listening proficiency in the investigation of the effect of the test-takers' L1 variable in paired speaking tests.

Following this review of some of the studies on the importance of listening in interactive speaking contexts, the next section will review some studies on the impact of L1 on L2 listening comprehension.

2.5 Impact of L1 on L2 listening comprehension

It is said that when language learners speak the L2, their L1 tends to transfer phonological features, e.g., pronunciation and accent of L2 (Harding, 2012). One of the most prominent aspects of L2 spoken language is accent (Isaacs, 2014), and the language learners can understand the L2 spoken by a speaker who shares the L1 background with them better than that spoken by a speaker from a non-shared L1 background or a NS (Bent and Bradlow, 2003). Harding (2012, p. 165) suggests that

> from the perspective of cross-language speech perception, there is a theoretical foundation for a shared-L1 effect based on the principle that L2 accents are primarily characterized by transfer from the L1, and those listeners who share a speaker's L1 will have an intimate familiarity with the phonological patterns of that speaker's L2 accent.

Buck (2001) argues that accent is potentially a very important variable in listening comprehension since when listeners hear an unfamiliar accent, it can cause a problem through which the whole comprehension process may be disrupted. Porter (1986, p. 209) states that

> where learners were from the same language background, pronunciation errors did not make language incomprehensible to other learners: virtually no breakdowns of communication occurred in learner-learner discussions owing to phonological problems; in the native-learner pairs, on the other hand, native speakers occasionally had trouble understanding learner phonology, as evidenced by clarification requests. The point is that learners from the same native language background may serve as better interaction partners for each other than learners from different language backgrounds on the basis that their similar interlanguage phonologies will be comprehensible.

Therefore, the non-shared L1 test-taker pairs may gain some disadvantages in interaction compared to the shared L1 test-taker pairs because a different L1 accent may cause difficulty in understanding each other properly. However, if the non-shared L1 test-takers are familiar with their partner's accent, the accent might not be problematic for them regarding understand their partner's speech. Furthermore, familiarity with a specific accent can facilitate the listener's comprehension (Derwing and Munro, 1997; Gass and Varonis, 1984). Familiarity with an accent not only affects the test-takers' intelligibility but also the way in which raters rate

the test-takers' scores, especially in the pronunciation category. When raters are familiar with test-takers' L1 and accent, they may better comprehend the performance and/or give a higher pronunciation score (Carey et al. 2011; Fayer and Krasinski, 1987; Winke et al., 2012). Additionally, having the same L1 as NNS test-takers, NNS raters could also affect the test-takers' pronunciation scores. NNS raters could give higher L2 pronunciation scores to NNS test-takers from the same home country as theirs than to those from a different country (Carey et al., 2011).

Bent and Bradlow (2003, p. 1600) state that

> individuals from the same native language background who are in the process of acquiring a given target language all share an 'interlanguage' and second-language learners often report that the speech of a fellow non-native talker is easier to understand than the speech of a native talker.

Bent and Bradlow explored the effects of speakers' L1 background on the intelligibility of the talkers' speech for listeners with the same L1 and those with a different L1. Five native talkers from three different L1 backgrounds (Chinese and Korean and English) were recorded reading simple English sentences which included keywords. Native listeners of four main groups, English (N=21), Chinese (N=21), Korean (N=10) and a mixed group from various native language backgrounds (N=12), completed a sentence recognition task by listening to the recordings of each talker and completing the missing words. The main finding of Bent and Bradlow's study was that the intelligibility of each talker depended on the listener's language background. For the non-native speaking learners who had the same L1 background, a non-native talker's speech was more intelligible to non-native listeners than to native listeners. This is because those learners shared phonetic and phonological knowledge from their L1 background and their linguistic knowledge covered both their L1 and their target language, while the linguistic knowledge shared between native/non-native pairs was only their linguistic knowledge of the target language.

The notion that speech produced by someone who shares an L1 background with the listeners is more comprehensible than the speech of someone with a non-shared L1 background is supported by Kachi's (2004) study. Kachi (2004) examined evaluative reactions of native and non-native listeners to the intelligibility

of Japanese English. Eighty-four undergraduate and graduate students from various L1 backgrounds (American English, Hindi, Mandarin and Japanese) at an American university participated in the study. Candidates were asked to listen to extemporaneous speech produced by Japanese English speakers from three different proficiency levels and American English speakers, evaluate it on an evaluative reaction scale and complete an intelligibility test (a word-for-word dictation task). Thirteen of these participants were selected to be interviewed. The results indicated that Hindi listeners were the most negative towards Japanese English, while American listeners were the most positive about it. Reactions gained from Chinese and Japanese listeners were similar. Japanese English was more intelligible to Chinese and Japanese listeners than American English. Kachi's findings suggest that English spoken by someone with a shared L1 background of the listener is easier to understand than by someone with non-shared L1 background. Data gained from the L1 Japanese interview participants revealed that "not only that they were empathetic toward the Japanese speakers in the affective domain, but also they could 'better understand' Japanese English cognitively because they could infer what the speakers really wanted to say, based on their linguistic and cultural background knowledge" (ibid, p. 199). In addition, Kachi found that the most powerful predictor of the listeners' intelligibility scores was their English proficiency.

Major et al. (2002) investigated the influence of the listeners' L1 on their comprehension of native and non-native varieties of English. Four groups of 100 listeners who were from different L1 backgrounds (Chinese, Japanese, Spanish and standard American English) listened to English lectures presented by speakers with different L1s (Chinese, Japanese, Spanish and standard American English) and answered questions based on the lectures. The findings showed that only Spanish L1 listeners understood English lectures given by Spanish speakers better than those given by other L1 speakers. Chinese L2 listeners who were tested on their comprehension of a spoken passage found Chinese L2 talkers significantly more difficult to understand than native speakers and more difficult than L2 Japanese and Spanish speakers. It is surprising that the Chinese listeners had more difficulty in

understanding English spoken by Chinese speakers than that spoken by other L1 speakers. Furthermore, the Japanese listeners did not comprehend the English lecture given by the Japanese speaker any better than that given by any of the other L1 speakers. The findings of Major et al.'s (2002) study showed that listeners did not always comprehend English spoken by the speaker who had the same L1 as them. The results of Major et al.'s study in the Japanese case (Japanese listeners did not find any advantages in listening to the Japanese lecturer compared to listening to the other L1-speaking lecturers) contrasts with Kachi's (2004) study which found that the English spoken by Japanese NSs was more understandable for Japanese listeners than English spoken by a native English speaker.

Stibbard and Lee (2006) followed up Bent and Bradlow's (2003) study. The participants were 50 listeners from four L1 groups (10 Koreans, 10 Saudi Arabians, 10 native English and 20 other mixed L1s). They were asked to listen to sentences which included familiar keywords. The sentences were read aloud by five talkers (Koreans and Saudi Arabians with high and low proficiency levels, and a native English speaker). The students' intelligibility was measured from their ability to recognise the keywords. The non-native listeners had the most difficulty in listening to the non-native talkers with a low proficiency level and a different L1 background to theirs. The finding indicated that the L1 talkers with a low proficiency level might have caused an intelligibility problem for listeners who were unfamiliar with English spoken with the non-shared L1 talker's accent, while this was not problematic when they were listening to the non-shared L1 talkers who had a high proficiency level. However, the evidence to support the notion that an intelligibility benefit was created by the non-native talkers with a high proficiency level for the listeners with a different L1 background or that the non-native talkers with a high proficiency level were less intelligible than the native talkers to listeners from any L1 backgrounds was not strong (Stibbard and Lee, 2006, p. 433).

The intelligibility benefit of having non-native listeners from the same L1 background as the talkers was explored by Algethami and colleagues (2011). Algethami et al. (2011) examined whether Saudi Arabian L2 listeners understood Saudi Arabic-accented English better than native English listeners did. Listeners

were divided into two groups: Australian English speakers and Saudi Arabic-accented English speakers. Nineteen participants from each group were asked to listen to and paraphrase English sentences produced by two groups of Saudi Arabic speakers (high and low pronunciation proficiencies as the experimental groups) as well as Australian English speakers (as the control group). The results showed the Saudi listeners gained slightly higher intelligibility scores than the Australian listeners from both L2 speakers' groups (high and low pronunciation proficiencies); however, the difference between both groups of listeners was not statistically significant. Algethami et al. concluded that although non-native listeners more easily understood L2 spoken by L2 speakers from the same L1 background, the intelligibility benefit they gained was small or there was no intelligibility benefit.

In the field of language testing, Harding (2012) investigated whether listeners who shared their L1 with speakers had an advantage in an academic English listening test. Two hundred and twelve L2 participants from different L1 backgrounds took three versions of the University Test of English as a Second Language (UTESL) listening test, including recorded materials of three different accented speakers: Australian English, Japanese and Mandarin Chinese. The results showed that Japanese L1 listeners had an advantage in a small number of items on the test which were recorded by the Japanese-accented speaker, while Mandarin Chinese L1 listeners had an advantage across several items in the test which were recorded by the Mandarin Chinese L1 speaker. The results of Harding's study illustrated that the listeners' degree of advantage in understanding a speaker who shared an L1 with them varied and depended on their L1 background.

Ockey and French (2014) questioned the impact of various accents on L2 listening comprehension tests. They investigated the extent to which strength and familiarity of accent affected L2 listeners' comprehension. The strength of accent scale used in their study was developed based on salience and comprehensibility (Derwing and Munro, 2009). Nine participating speakers of English were selected based on a judgement of their accent strength. There was one American speaker (judged as an accent representative of the local variety), four Australian and four British English speakers. Participants were TOFEL iBT (Test of English as a

Foreign Language Internet-based Test) test-takers (N=21,726) from 148 countries. They were randomly assigned to listen to a common lecture, 686 words in length, on a natural sciences topic presented by one of nine speakers. While listening to the lecture, several context-relevant photographs of the speaker were shown on a computer screen. The test-takers were allowed to take notes while listening and used their notes when answering comprehension test questions. After they had finished listening to the lecture, they were asked to answer six listening comprehension questions and this was followed by a questionnaire response session. A negative relationship between strength of accent and listening test scores was found. That is, as the strength of an accent increased, the listening test scores decreased. However, this effect was significant only when the strength of accent score based on the strength of accent scale was stronger than two. The results of Ockey and French's study suggested that strong accents for which some listeners required extra effort for complete comprehension tended to obstruct comprehension and that listeners gained a comprehension benefit when listening to a familiarly accented speaker.

- The studies reviewed above provide evidence with which to argue that there seem to be some effects of English spoken by those with non-shared L1s on different L1 listeners' comprehension in some contexts. The speaker's proficiency seems to affect the listener's comprehension. For non-native listeners, the intelligibility of a non-native talker from a different L1 background and with a high proficiency level is greater than or equal to the intelligibility of a native talker (Bent and Bradlow, 2003, p. 1607). Non-native listeners who have the same L1 background as speakers are more able to comprehend the speakers' speech than listeners who have different L1 background with the speakers (Bent and Bradlow, 2003; Harding, 2012; Kachi, 2004). Their ability to infer what the speakers intend to say based on the linguistic and cultural background knowledge they shared seems to be helpful (Kachi, 2004). Different L1-related accents can cause difficulty in their L2 listening comprehension (Harding, 2012; Ockey and French, 2014; Stibbard and Lee, 2006). Cross-linguistic influence on L2 pronunciation is a powerful factor which affects the interaction. For example, a conversation between Thai and English native speakers, Thai speaker pronounced each word

with ed equal weight regardless of the information structure of the discourse and produced the discourse with gradual downward intonation tended to trigger troublesome occurrence in the conversation (Wennerstrom, 2000). The low pitch used in English indicates the end of thought and signals the listener to take the floor. When it is exploited inaccurately, it results an unwanted interruption from the native listener which possibly impedes the conversation (Wennerstrom, 2000).

Since many studies suggest that the L1 backgrounds of the speakers and the listeners could affect L2 listening comprehension, it is essential to investigate whether non-native test-takers' L1 background affects their L2 interactive listening and their speaking performance in paired speaking test formats. Moreover, research has shown that raters' familiarity with the test-takers' accent could affect their rating (Carey et al., 2011; Fayer and Krasinski, 1987; Winke et al., 2012). In the present study, in order to avoid rater familiarity being a confounding variable, raters who are not familiar with the selected L1s have been chosen. There has been no study, to the researcher's knowledge, which has systematically looked at the effects of L1 on test-takers' listening comprehension in paired speaking tests. Therefore, this study fills this gap by examining both test-takers' speaking and listening proficiencies and L1 factors in the paired formats and identifying interactive listening-related communication patterns in shared L1 pairs and non-shared L1 pairs.

2.6 Summary of literature review

This chapter has reviewed some of the literature related to development and concerns of utilising paired speaking assessments. Firstly, the theory base for the study was presented, detailing Weir's (2005; further elaborated in Taylor, 2011) socio-cognitive framework for validating a speaking test. Of particular relevance to this study are three of the components involved in this framework: test-taker characteristics, cognitive validity and context validity, and the subsequent sections of this chapter reviewed the relevant variables involved in each component. Secondly, the development of paired speaking assessment and issues related to the

OPI were described. Thirdly, factors which could affect the co-construct of paired and group interactions were explained. Fourthly, a listening comprehension model (Field, 2008) and the importance of interactive listening in interaction were described. Finally, impacts of L1 on L2 listening comprehension were discussed. The reviewed literature suggests that L1 characteristics of test-takers and interlocutors affect their speaking performance in paired formats. Despite English being widely used as an international language, to date there have been only a handful of studies related to L1 features in NNS–NNS interaction (e.g., Isaacs, 2013; Jenkins, 2002; Lu, 2010). Moreover, to the researcher's knowledge, no research has been systematically conducted to investigate whether non-native speaking test-takers from the same and different L1 backgrounds gain an advantage or a disadvantage because of their L1 background during the paired interaction and how it relates to their listening proficiency. This study, therefore, aims to fill these gaps by examining the impact of test-takers' L1 backgrounds and listening proficiency on paired test scores and discourse features.

CHAPTER 3 Research Methodology

This chapter describes the methodological approach used in the present study. The research questions and research design are presented first. Data collection and the analysis methods of the pilot study and its results are then illustrated. Finally, it presents the methodology of the main study, while highlighting how it has been improved in light of the pilot study experience.

3.1 Research questions

This research addresses three research questions related to test-takers' listening proficiency, paired speaking performance and L1 background variables. The research questions are as follows:

RQ1: To what extent is test-takers' performance in paired speaking tests in shared and non-shared L1 pairs affected by their listening proficiency?

RQ2: Are there any differences in paired speaking scores when test-takers are paired with shared L1 partners as compared to (when they are paired with) non-shared L1 partners?

RQ3: What are the similarities and differences in communication patterns between shared and non-shared L1 pairs?

A number of studies reviewed in Chapter 2, such as those of Galaczi (2004, 2013), Nakatsuhara (2012), Ducasse (2010) and Ducasse and Brown (2009), provide evidence to support the importance of listening to achieve successful spoken interaction in interactive speaking tests. RQ1 therefore aims to verify whether/to what extent test-takers' L2 speaking performance in shared and non-shared L1 pairs is affected by their L2 listening proficiency.

To achieve successful speaking interaction, participants need to actively act as both a speaker and a listener. As reviewed in Chapter 2, learners' L1 characteristics, including phonologic features, tend to transfer to their L2 English (Harding, 2012; Jenkins, 1995, 2000). When L2 test-takers from the same L1

background communicate with each other in L2, they may understand each other better than those who are from a different L1 background (e.g., Bent and Bradlow, 2003; Harding, 2012; Jenkins, 1995, 2000; Kachi, 2004; Porter, 1986). RQ2 will address this issue by examining whether test-takers perform differently in shared and non-shared L1 pairs. The review of the literature has suggested that different cultural and L1 backgrounds may affect test-takers' use of a discourse feature (e.g., Lu, 2010). RQ3 therefore intends to examine communication patterns between shared and non-shared L1 pairs in relation to communicative effectiveness and interactive listening behaviour during paired tests.

3.2 Research design

To answer the above research questions, this study employed a mixed methods approach which combined both qualitative and quantitative approaches. Creswell and Plano Clark (2011) and Dörnyei (2007) state that mixed methods research provides more comprehensive results than either qualitative or quantitative research alone, and it provides strengths that compensate for the weakness of both types of research. The weakness of quantitative research is a lack of understanding of detailed study contexts and direct consideration of participants' voices, while qualitative research is weak regarding its power to generalise findings to a large population and its tendency to suffer because of personal biases and interpretations of researchers.

With the hope of obtaining greater insight into the relationship between test-takers' L1, their listening proficiency and their speaking performance in pairs, the mixed methods approach was utilised. The quantitative approach was used to analyse the questionnaire responses, correlations between listening and various speaking test scores (i.e., monologic and paired speaking scores with shared and non-shared L1 partners), differences between monologic and paired speaking test scores, and differences in paired speaking scores under two test conditions (with shared and non-shared L1 partners). More details will be presented in Sections 3.3.5.1 and 3.3.5.2.

Interactional data gained from video recordings of test-takers' speaking performance in pairs was transcribed following Atkinson and Heritage's (1984)

transcription symbols (see Appendix 6) and analysed by using CA methods (see Glossary and Section 3.3.4.2). CA was employed to explore the test-takers' communication patterns related to interactive listening. A coding scheme was developed to analyse data from stimulated recall interviews with test-takers and raters (see Sections 3.3.2.4 and 3.4.2.2), and findings from interviews were used to triangulate the CA results. It should be pointed out that there was no stimulated recall interview with raters in the pilot study.

Among various models of mixed methods research, the study used a convergent parallel mixed methods design (Creswell and Plano Clark, 2011), where quantitative and qualitative data were collected in two parallel strands, analysed separately and then the findings were integrated. The two data strands provide different types of information and allow for an in-depth and comprehensive set of findings. Figure 3.1 presents information on the data collection and analysis strands in the research design.

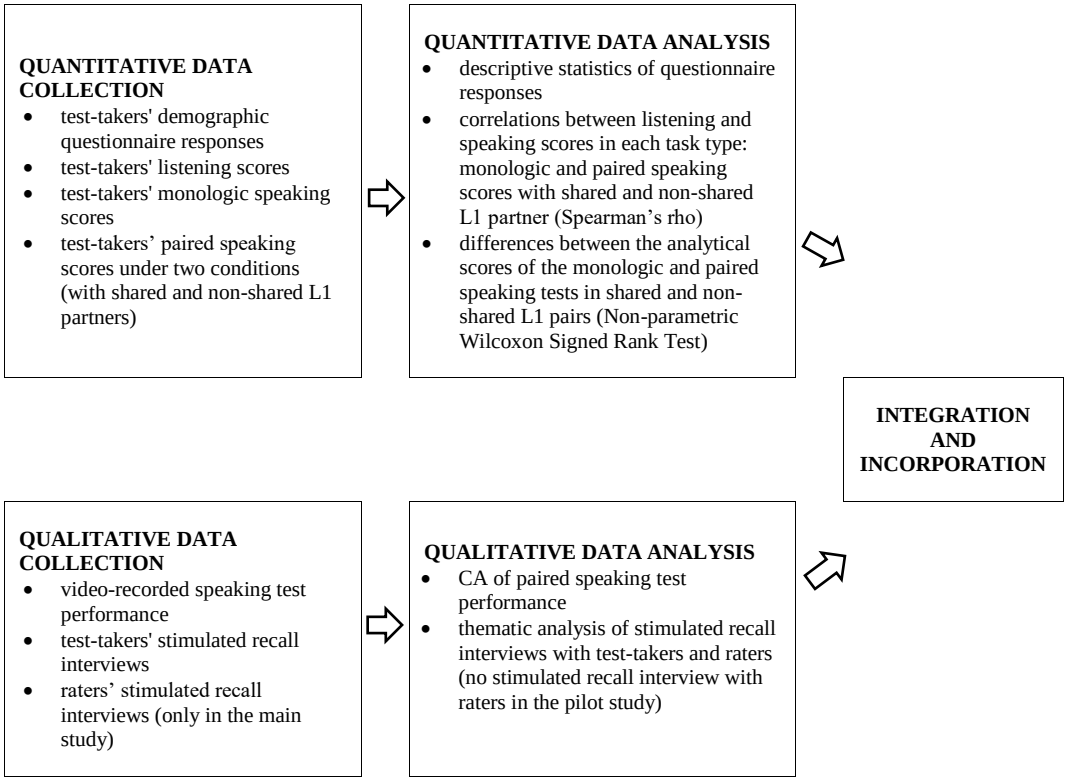


Figure 3.1: Framework for research design

3.3 Pilot study

In light of the literature review in Chapter 2, a small-scale pilot study (N=12) was carried out to verify or modify the research instruments for the main study. The participants and research instruments of the pilot study and their details will be presented in Sections 3.3.1 and 3.3.2.

3.3.1 Participants in the pilot study

Twelve test-takers and two raters participated in the pilot study.

3.3.1.1 Test-takers in the pilot study

Twelve students who were taking a pre-sessional English language programme at the University of Bedfordshire were the participants in the pilot study. In particular, students with Thai and Gujarati (1st round of data collection) and Thai and Arabic (2nd round of data collection) L1 backgrounds were selected by employing a purposive sampling method, where the selection of samples is made by the researcher based on who are the most suitable representatives of the issues involved in the research (Coolican, 2013); in this case, these were participants from different L1 backgrounds with a similar English speaking and listening proficiency based on their IELTS scores. IELTS was thought to be the most appropriate tool to inform the participants' English proficiency in this study, as all pre-sessional programme students had recent IELTS scores certifying their English proficiency so that they could be admitted to universities in the UK.

As detailed in Table 3.1, data was collected in two rounds from 12 test-takers in total: six Thai, two Gujarati and four Arabic L1 speakers. The first round involved two Thai and two Gujarati L1 test-takers. However, due to limited availability of Arabic L1 students, the second round of data was collected from four Arabic L1 speakers (instead of Gujarati L1 speakers) from Lebanon, Libya and Egypt, and four Thai speakers.

Table 3.1: Test-takers' L1, country of origin, gender, age, IELTS speaking and listening scores and pairing - pilot study

Pilot study round	ID	L1	Country of origin	Gender	Age	IELTS examination score		Type of pair	
						S	L	SL	NSL
1	I1	Gujarati	India	Male	22	5.5	5.5	I1I2	I1T5
1	I2	Gujarati	India	Male	23	5.5	5.5		
1	T5	Thai	Thailand	Female	35	5.0	5.0	T5T6	I2T6
1	T6	Thai	Thailand	Female	42	5.0	5.0		
2	A1	Arabic	Lebanon	Male	33	5.0	5.0	A1A2	A1T1
2	A2	Arabic	Egypt	Male	29	5.5	5.0		
2	T1	Thai	Thailand	Female	38	5.5	6.0	T1T2	A2T2
2	T2	Thai	Thailand	Female	27	6.0	6.0		
2	A3	Arabic	Libya	Female	24	5.0	5.0	A3A4	A3T3
2	A4	Arabic	Egypt	Female	28	5.5	5.0		
2	T3	Thai	Thailand	Male	28	5.5	5.5	T3T4	A4T4
2	T4	Thai	Thailand	Female	24	5.5	5.5		

Note: S refers to speaking proficiency. L refers to listening proficiency.
SL refers to shared L1 pairs. NSL refers to non-shared L1 pairs.

The test-takers were, on average, 29 years old (Mean=29.42, SD=6.33). The youngest was 22 and the oldest was 42 years old. The length of their stay in the UK ranged from two months to 26 months (Mean=12.83, SD=10.50). According to their IELTS scores, all test-takers had similar English speaking and listening proficiency. Their speaking and listening scores ranged from Bands 5.0 to 6.0. A demographic questionnaire (see Appendix 1) also included Likert-scale questions regarding test-takers' perception of their own familiarity with English spoken by shared and non-shared L1 speakers. All test-takers reported their familiarity with English spoken by shared L1 speakers. However, regarding their familiarity with English spoken by non-shared L1 speakers, while all Gujarati and Arabic L1 speakers selected 'neutral', i.e., neither familiar nor unfamiliar, for the English spoken by Thais, Thai L1 participants expressed more varied views towards the English spoken by Arabic or Gujarati L1 speakers; of the two paired with a Gujarati speaker, one reported 'neither familiar nor unfamiliar' and one 'unfamiliar'; of the four with an Arabic speaker, one chose 'very familiar', one 'familiar', one 'neither familiar nor unfamiliar', and one 'unfamiliar'.

The test-takers' gender, length of stay in the UK, differing amounts of exposure to the English language inside and outside the classroom, and familiarity with the English spoken by the shared and non-shared L1 speakers were not

controlled in the pilot study, and this issue will be revisited when the design of the main study is presented.

3.3.1.2 Speaking test raters

Raters' scores could be influenced by whether they share a common language background with the test-taker (Hamp-Lyons and Davies, 2008; Kenkel and Tucker, 1989). Therefore, English native speakers who are not familiar with Thai, Gujarati or Arabic were asked to participate in the pilot study. Two British female raters with experience in spoken language assessment were involved.

Rater 1 has a Diploma in Teaching English as a Foreign Language and Teaching English as a Second Language, and has had 20 years' experience in teaching English to international students and 15 years' experience in examining speaking skills at Trinity College, London. She is a native speaker of English but also speaks Spanish, French, Italian and a little German, but it was confirmed that she is not familiar with the three languages selected for the pilot study.

Rater 2 is also unfamiliar with the three languages, although she speaks Italian, Spanish and Portuguese in addition to her native language, English. She holds Diplomas in CELTA (Certificate in English Language Teaching to Adults) and in DELTA (Diploma in English Language Teaching to Adults). She has a bachelor's degree in the Spanish Language and Third World Studies, and has a master's degree in Applied Linguistics. She has had 13 years' experience in teaching English to international students and six years' experience in examining speaking for IELTS.

Although they were experienced examiners/raters for standardised speaking tests, it has to be pointed out that no rater training was provided for the FCE test used in the pilot study. This was a possible limitation of the pilot study, which was remedied in the main study.

3.3.2 Research instruments of the pilot study

The pilot study aimed to try out the main study instruments for investigating the relationship between test-takers' L1 and their listening and speaking proficiencies. This study shares some similarities with Nakatsuhara's (2011) study, which examined the relationship between test-takers' listening proficiency and their

performance in the IELTS Speaking Test. This study therefore incorporated some of Nakatsuhara's research instruments and data collection methodologies. The research instruments used in the pilot study consisted of the following:

- a background questionnaire;
- a listening test;
- speaking tests (a monologic task and two paired tasks); and
- stimulated recall interviews with test-takers.

The following section explains these research instruments in detail.

3.3.2.1 Questionnaire

The background questionnaire (see Appendix 1 for a full questionnaire) was used to gather the following information from test-takers:

- test-takers' demographic information, such as gender, age, country of origin, length of stay in the UK;
- test-takers' English test scores (overall, speaking and listening) in a standardised examination (i.e., IELTS) that they have recently taken; and
- test-takers' opinion of their own familiarity with English spoken by shared and non-shared L1 speakers. Likert-scale questions were used in this section.

3.3.2.2 Listening test for the pilot study

A listening test from the Cambridge First Certificate in English (FCE) Examination was considered to be appropriate to measure the test-takers' listening proficiency in the pilot test. As detailed in Section 3.3.1.1 above, test-takers were recruited from a university's pre-sessional course, and were assumed to have B1–B2 levels of English. The FCE is a B2 test, and gaining the FCE is a prerequisite for entrance to some UK universities and other educational institutes (Cambridge ESOL, 2009, p. 4). In addition, the purpose of the test was congruent with the objective of this study, i.e., measuring listening required for spoken interaction, since the FCE listening test aims to test students' general English listening proficiency rather than academic listening ability. The items in the listening test for the pilot study were therefore taken from FCE practice material published by the University of Cambridge ESOL Examinations (Cambridge ESOL, 2009). The test included 30 items in total, each of the 30 items being worth one mark. It consisted of multiple-choice, sentence-completion and multiple-matching tasks (see Appendix 2 for the

listening test of the pilot study). The structure of the listening test is presented in Table 3.2.

Table 3.2: Structure of a listening test (40-minute test) part of the FCE Examination (UCLES, 2009)

Part	Task type and focus	Format	No. of items
1	Multiple choice. General gist, detail, function, purpose, attitude, opinion, relationship, topic, place, situation, genre, agreement, etc.	A series of short unrelated extracts, of approximately 30 seconds each, from monologues or exchanges between interacting speakers. There is one multiple choice question per text, each with three options.	8
2	Sentence completion. Detail, specific information, stated opinion.	A monologue or text involving interacting speakers and lasting approximately 3 minutes. Candidates are required to complete the sentences with information heard on the recording.	10
3	Multiple matching. General gist, detail, function, purpose, attitude, opinion, relationship, topic, place, situation, genre, agreement, etc. General gist, detail, function, purpose, attitude, opinion, relationship, topic, place, situation, genre, agreement, etc.	Five short related monologues, of approximately 30 seconds each. The multiple-matching questions require selection of the correct option from a list of six.	5
4	Multiple choice. Opinion, attitude, gist, main idea, specific information.	A monologue or text involving interacting speakers and lasting approximately 3 minutes. There are seven multiple-choice questions, each with three options. A monologue or text involving interacting speakers and lasting approximately 3 minutes. There are seven multiple-choice questions, each with three options.	7
	Total		**30**

The listening test was administered in line with the Cambridge FCE test protocol; the candidates listened to each listening test part twice and the test took approximately 40 minutes in total under exam conditions. The tests were marked according to the answer keys in the practice materials (Cambridge ESOL, 2009).

3.3.2.3 Speaking test for the pilot study

Three speaking tasks were administered in the pilot study: a monologic task and two paired tasks (task A and task B). Part 3 collaborative tasks in the Cambridge

FCE speaking test, taken from practice materials published by the University of Cambridge ESOL Examinations (Cambridge ESOL, 2009), were used for paired tasks in this study.

The FCE Speaking Test was again thought to appropriately target the proficiency level of participants in the pilot study. The FCE is equivalent to Level B2 on the Common European Framework of Reference (CEFR) scale, which is required for using English in "*everyday written and spoken situations, and achieving a certificate at this level that a candidate is becoming skilled in English*" (ibid, p. 3). The purpose of the FCE Speaking Test is to assess test-takers' ability to communicate effectively in face-to-face situations (UCLES, 2015). The total time for the FCE Speaking Test is approximately 14 minutes per pair of candidates. The FCE Speaking Paper comprises four tasks: *interview*, *individual long turn*, *collaborative* and *discussion tasks*. Details of each task are presented in Table 3.3.

Table 3.3: Structure and tasks of the FCE Speaking Test (UCLES, 2015, p. 71)

Part 1:	**Interview**
Task type and format:	A conversation between the interlocutor and each candidate (spoken questions).
Focus:	The focus is on general interactional and social language.
Timing:	2 minutes
Part 2:	**Individual long turn**
Task type and format:	An individual 'long turn' for each candidate with a response from the second candidate. In turn, the candidates are given a pair of photographs to talk about.
Focus:	The focus is on organising a larger unit of discourse; comparing, describing and expressing opinions.
Timing:	A 1-minute 'long turn' for each candidate, plus a 30-second response from the second candidate. The total time for Part 2 is 4 minutes.
Part 3:	**Collaborative task**
Task type and format:	A two-way conversation between the candidates. The candidates are given spoken instructions with written stimuli, which are used in discussion and decision-making tasks.
Focus:	The focus is on sustaining an interaction; exchanging ideas, expressing and justifying opinions, agreeing and/or disagreeing, suggesting, speculating, evaluating, reaching a decision through negotiation, etc.
Timing:	A 2-minute discussion followed by a 1-minute decision-making task. The total time for Part 3 is 4 minutes.
Part 4:	**Discussion**
Task type and format:	A discussion on topics related to the collaborative task (spoken questions).
Focus:	The focus is on expressing and justifying opinions, agreeing and/or disagreeing and speculating.
Timing:	The total time for Part 4 is 4 minutes.

Of relevance to this study is Part 3, the collaborative task. It aims to assess "*the candidates' ability to engage in a discussion and to work towards a negotiated outcome of the task set*" (UCLES, 2015, p. 72). The candidates are requested to respond to written and visual stimuli, and are expected to discuss the task prompts with their partner within 4 minutes. The task focuses on language functions such as "*sustaining an interaction, exchanging ideas, expressing and justifying opinions, agreeing and/or disagreeing, suggestion, speculating, evaluating, reaching a decision through negotiation*" (Cambridge ESOL, 2009).

In this study, the paired speaking tests were employed to assess candidates' interactive speaking proficiency potentially involving listening proficiency. Two versions of the paired speaking task were prepared so that one would be used for the shared L1 condition and the other one for the non-shared L1 condition. The two paired speaking tasks consisted of two questions and several pictures, as shown in Appendix 4. The questions were as follows:

Task A: How difficult is it to be successful in these professions?

In which profession is it most difficult to get to the top?; and

Task B: What are the advantages of having friends?

In which situation are friends most important?

Each task was printed on a card measuring 1.5 feet x 2 feet. They were shown on a table, and each pair shared the same card when performing their paired speaking test. The order of the two paired speaking task prompts was counterbalanced to control for a potential prompt effect. The order of pairings (shared and non-shared L1 pairs) in the two test sessions were also counterbalanced. Each pair was asked to discuss the given topic for 4 minutes. To make sure that the participants understand what was expected in the paired speaking tasks, a brief explanation of the task requirements was provided prior to the tests. Following the current FCE practice, planning time was not allowed. The administration of the paired speaking test will be presented later in Section 3.3.3.

The monologic speaking task was also developed from the FCE Part 3 collaborative task, rather than using a monologic task from the FCE Part 2 individual long turn. This is because the monologic task in this study had to be as equivalent as possible to the paired speaking tasks in terms of topical and linguistic

demands by selecting questions which required test-takers to provide reasons to support their opinion, in order to enable comparison of test-taker performance in the two tasks. The original FCE Part 3 collaborative task was therefore modified in a way that meant that test-takers were required to respond to two questions on their own for 2 minutes, instead of discussing them with their partners. The two questions (i.e., *How important are these things for a happy life? Which two are the most important?*) were accompanied by seven pictures (see Appendix 3). This adaptation allowed raters to assess test-takers' monologic speaking proficiency without any interference from test-takers' listening proficiency or their partner's contribution. Only one version of the monologic speaking task was developed, and it was used with all the test-takers. To make the task condition similar to the paired tasks, no planning time was allowed prior to the monologic task. The administration of the monologic speaking test will be described later in Section 3.3.3.

3.3.2.4 Stimulated recall with test-takers

Stimulated recall is a technique which can be used to gain insights into participants' cognitive processes, thoughts or feelings they had while performing a test task or activity. It is usually carried out immediately after participants have finished a task or activity by utilising audio recordings or video recordings of their performance as a stimulated recall tool (Gass and Mackey, 2000). Since this study aimed to gain insights into the test-taker interaction in shared and non-shared L1 pairs, a stimulated recall interview was thought to be appropriate to obtain greater understanding of test-takers' interactive communication skills related to interactive listening during the interaction.

In the pilot study, the stimulated recall interview was conducted with all test-takers. Each test-taker was interviewed immediately after finishing each paired speaking task. A video recording of the test-takers' paired speaking performance was used to stimulate their memory while performing the paired speaking test. Examples of questions used in the stimulated recall interview were as follows:

- Why did you say X?
- I saw you frowned while listening to your partner. What were you thinking? Did you understand what your partner was saying?
- What were you thinking while your partner was speaking for quite a long time?

- What were you thinking when pronouncing "*uh huh*", "*mm*", "*yeah*", "*yeah yeah*"?
- You said a word in your L1. What were you thinking at that time?
- Why did not you answer your partner's question? What were you thinking at that time?

The details of stimulated recall interview procedures are presented in Section 3.3.3. All stimulated recall interviews were audio recorded and then transcribed orthographically by the researcher.

3.3.2.5 Raters' scores in monologic and paired speaking tests

An operational FCE speaking test session involves two examiners. One examiner takes the role of interlocutor and is responsible for managing the test and providing a global assessment for each candidate's performance, while the other takes the role of assessor and does not participate in the interaction but focuses on awarding each candidate analytical scores by using four analytical categories: *grammar and vocabulary, discourse management, pronunciation* and *interactive communication* (UCLES, 2009, p. 81). The candidates are assessed based on their own performance, rather than a joint performance. Although some researchers like May (2009) advocate joint rating of interactive communication, this study followed the current FCE rating practice of rating each candidate separately. As reviewed in Chapter 2, this individual scoring seems to ensure fairness to test-takers in a situation in which a test-taker refuses to collaborate despite the partner's invitation to collaborate (e.g., Nakatsuhara, 2013).

The operational test assesses test-takers' overall performance across the four parts of the test, but in order to compare scores from different tasks, the raters in this study were asked to rate each of the paired and monologic tasks separately based on the four rating categories:

- **Grammar and Vocabulary** refers to the ability to use grammatical forms and vocabulary accurately and appropriately. Performance is assessed through the overall effectiveness in using language in spoken interaction.
- **Discourse Management** refers to the ability to link utterances by forming coherent speech without improper hesitation, and to produce utterances relevant to the tasks and logically to develop the themes or arguments of the tasks.
- **Pronunciation** refers to the ability to produce intelligible utterances to achieve the requirements of the task including stress, intonation and individual sounds.

- **Interactive Communication** refers to the ability to develop the discourse actively, participate in the range of interactive situation and develop discussions on the topic range by initiating and responding appropriately as well as properly deploying strategies to maintain interaction and fulfil the task (UCLES, 2009, p. 86).

Additionally, unlike the operational test, the speaking tests in this study were administered by the researcher, and non-live rating was carried out. All test-takers' speaking performances in the monologic and paired tests were video recorded and the files were copied onto DVDs. The DVDs were sent to two raters to individually assess the test-takers' speaking performance. Special care was taken regarding the order of speaking test recordings on the DVDs. The order of all recordings was carefully arranged to prevent the previous performance of the same candidate from influencing raters' judgement of the same candidate's performance on another test. The raters used the video recordings to rate the test-takers' speaking ability, employing a publicly available version of the Cambridge FCE rating scale (UCLES, 2009; see Appendix 5 for the rating scale). The band score for each category ranged from 0 to 5 and the raters were allowed to award scores in 0.5 increments. Since the interactive communication category is not relevant for the monologic speaking test, it was not used in rating the monologic speaking performance. Therefore, the total possible score for the monologic speaking test was 15 and the total possible score for each paired speaking test was 20. Since there were two raters scoring the test-takers' speaking performance, the scores used for quantitative analysis were average scores gained from the two raters.

3.3.3 Data collection procedures of the pilot study

Staff from the Language Centre at the University of Bedfordshire were informed via email about the research and asked to help contact lecturers who taught students in the pre-sessional English language programme. After receiving a response from the Language Centre staff, the researcher sent an email to those lecturers to ask for permission to gather data from their students. Upon receiving their permission, the researcher went to their classes to meet students and ask for volunteers for the research.

After gaining students' background data, two rounds of speaking data collection were carried out with participants from three L1 backgrounds: Gujarati,

Arabic and Thai. As noted earlier, in Section 3.3.1.1, the first round was with two Thai L1 and two Gujarati L1 students, and the second round was with four Thai L1 and four Arabic L1 students.

Before starting the data collection, test-takers were informed about the research objectives and asked to sign consent forms to confirm that they wished to participate in the study. All test-takers were asked to complete the background questionnaire first and subsequently completed the listening test.

The listening test was administered by the researcher in a quiet room following standard Cambridge FCE listening test practices. The instructions for each listening test task were given on the question paper, and were also heard on the recording. The listening test took approximately 40 minutes.

Test-takers were asked to do the monologic speaking test after finishing the listening test. The test-takers completed the monologic test task one by one. During the monologic speaking test, only the test-taker and the researcher were in the room. The test-taker was given the instructions both orally and in a written format before performing the task. The test-taker was then given the monologic task prompt. As noted in Section 3.3.2.3, there was no planning time for the test-taker before the monologic speaking performance. The test-taker was first of all asked to introduce himself/herself to the researcher briefly as a warm-up task, and was then asked to do a monologic task for 2 minutes. The test-taker was told to stop speaking when the time was finished. The next test-taker was then invited to do the test.

When all test-takers had finished the monologic speaking test, they were asked to do the paired speaking test. Although the researcher wanted to counterbalance the order of monologic and paired tests, this was not possible due to practical constraints. The test-takers took two paired speaking test tasks, one with a shared L1 partner and one with a non-shared L1 partner, in a counterbalanced manner. During the test, only two test-takers and the researcher were in the room. Following the current FCE practice, no planning time was allowed prior to the paired speaking test. The test-takers were given instructions orally and in a written format before performing their speaking interaction. As a warm-up task, they were asked to introduce themselves to each other briefly, and then they were given a paired task prompt and asked to start the task for 2 minutes. The test-takers were told to stop

speaking after 4 minutes. Both the monologic and the paired test sessions were video recorded and audio recorded. The audio recordings were just a back-up in case the video recordings' quality was poor.

After each pair finished their speaking performance, one test-taker was asked to go outside the room. The test-taker, who was inside the room, was interviewed by using a stimulated recall approach. Before the interview, the test-taker was given the instructions both in oral and written formats about how to provide verbal reports. The video recording of the test-taker's performance in the paired task was shown to stimulate the test-taker's memory. Either the test-taker or the researcher could stop the video. When the test-taker could remember what he/she was thinking at the time he/she was talking to his/her partner, he/she told the researcher to pause the video. The video was also paused when the researcher wanted to ask the test-taker something related to the interactive communication. Examples of question used in the stimulated recall interview were presented earlier, in Section 3.3.2.4. The interview was audio recorded. When the interview was over, the test-taker was asked to go outside the room and his/her partner was invited to be interviewed next. When the interview of the first pair was finished, the next pair was asked to do the test. This process of the paired speaking test was repeated with all the pairs.

3.3.4 Data analysis of the pilot study

Details of score analysis as well as interactional data analysis are set out in this section.

3.3.4.1 Score analysis

Scores from the listening test and monologic and two paired speaking tests were statistically analysed using the SPSS program. The strength of the correlations between the listening scores and the analytical scores of monologic and paired speaking tests in two types of pairs was examined by using Spearman correlation to answer *RQ1*. Spearman correlation is used to "correlate data when it is ordinal, when data is not normally distributed" (Hinton et al., 2004, p. 300) and was thought to be appropriate to apply to the small data set of the pilot study. To answer *RQ2*, a non-parametric Wilcoxon Signed Rank Test was used to examine the difference

between the analytical scores of paired speaking tests in shared and non-shared L1 pairs. A non-parametric Wilcoxon Signed Rank Test, which is an ordinal-level difference test, used to investigate the differences between two related data sets when the score data are not normally distributed (Coolican, 2009), was thought to be suitable for the small-scale data set of the pilot study.

3.3.4.2 Interactional data analysis

As noted in Section 3.2, although quantitative analysis is beneficial in that it provides generalisation of the findings, its limitation is that it does not provide in-depth information about what is really happening in the interaction. Hence, qualitative CA analysis was utilised, together with quantitative analysis, in order to a gain better understanding of interaction between shared and non-shared L1 pairs. According to Hutchby and Wooffitt (1998, p. 14), CA is

> the study of recorded, naturally occurring talk-in-interaction with the aim to discover how participants understand and respond to one another in their turns at talk, with a central focus being on how sequences of actions are generated ... to uncover the tacit reasoning procedures and sociolinguistic competencies underlying the production and interpretation of talk in organised sequences of interaction.

Ten Have (1999, p. 102) notes that "CA involves both an 'inductive' search for patterns of interaction, and an explication of the emic logic that provides for their significance." Therefore, CA is concerned with how participants understand and respond to each other in the interaction. The CA transcription convention advocated by Atkinson and Heritage (1984) was used for the analysis. The transcription symbols are presented in Appendix 6.

CA was carried out on paired test data to investigate communication patterns in paired speaking tests which might be related to test-takers' listening ability and their L1 background factors (to answer *RQ3*). In addition, CA was used to discover any similarities and differences in communication patterns related to interactive listening between shared and non-shared L1 pairs. The researcher's interpretation of communication problems presented in CA analysis was supplemented by test-takers' stimulated recall interview data in order to triangulate CA results to achieve greater reliability of the qualitative analysis.

3.3.5 Results of the pilot study

This section presents the results of the pilot study. It should be noted that these results can only be considered as suggestive on account of the limited number of the participants (N=12) involved in the pilot study.

3.3.5.1 Relationship between listening and speaking scores in monologic and paired speaking tests (RQ1)

Descriptive statistics and a histogram of students' listening scores are presented in Table 3.4 and Figure 3.2. Although the test was thought to target the participants, they did not perform as well as expected. The highest score for the listening test was 18 and the lowest was 5 out of 30. On average, they scored 12.50. Due to the small sample size, it was not possible to examine the reliability of the listening test.

Table 3.4: Statistics of test-takers' listening scores (N=12) – pilot study

	Min	Max	Mean	SD
Listening test (30 items)	5.00	18.00	12.50	3.58

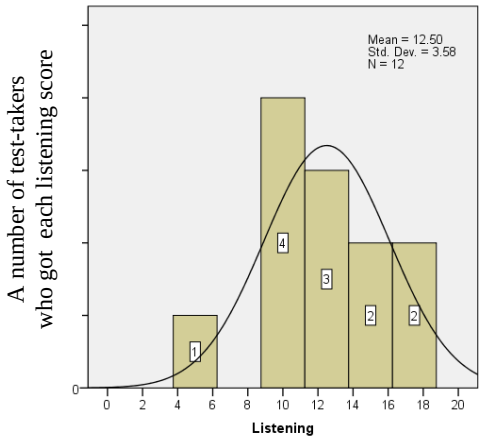


Figure 3.2: Histogram for test-takers' listening scores – pilot study

For the speaking test scores given by the two raters, the inter-rater reliability of the two raters was examined by absolute agreement rates. Among all scores given by the two raters, only 20.24% showed the exact agreement, and 83.33% showed the exact and adjacent agreement (a difference of 0 or 1 point; more details are reported in Jaiyote, 2015). Although reasonable inter-rater reliability seemed to be attained, it did not reach an acceptable agreement level (90% for the combination of the exact and adjacent agreement) as suggested by Graham et al. (2012). This issue will be

revisited in Section 3.3.6. When there were discrepancies between the two raters' scores, average scores were awarded to test-takers.

Descriptive statistics and histograms of students' speaking scores are presented in Table 3.5 and Figures 3.3–3.9.

Table 3.5: Statistics of test-takers' analytical scores in the monologic and paired speaking tests (N=12) – pilot study

Category	Mode	Mean	Median	SD	Min	Max
Grammar and Vocabulary (1-5 points)	**Mono**	2.54	2.38	.68	1.75	4.00
	Paired	2.74	2.76	.39	2.13	3.25
Discourse Management (1–5 points)	**Mono**	2.90	2.75	.69	1.75	4.00
	Paired	2.92	2.94	.53	2.13	3.63
Pronunciation (1–5 points)	**Mono**	2.42	2.50	.64	1.50	3.25
	Paired	2.69	2.50	.52	1.63	3.50
Interactive Communication (1–5 points)	**Mono**	-	-	-	-	-
	Paired	3.02	3.07	.57	2.13	3.75

Note: Total possible score for each category is 5 points

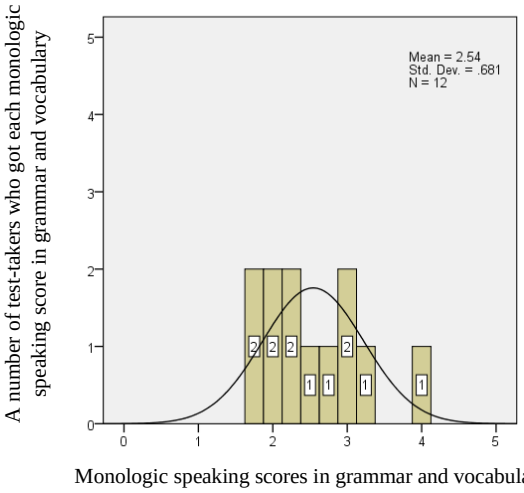


Figure 3.3: Histogram for monologic speaking scores in grammar and vocabulary – pilot study

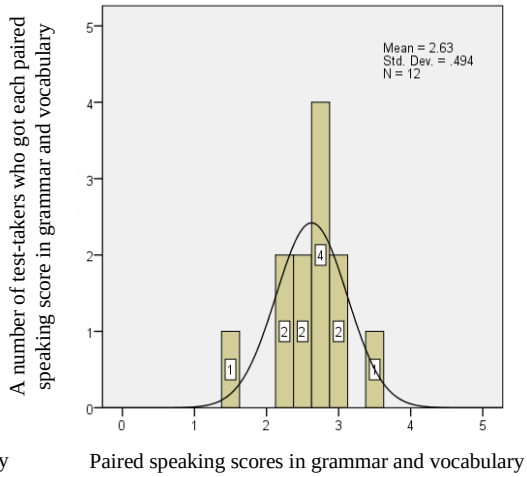


Figure 3.4: Histogram for paired speaking scores in grammar and vocabulary – pilot study

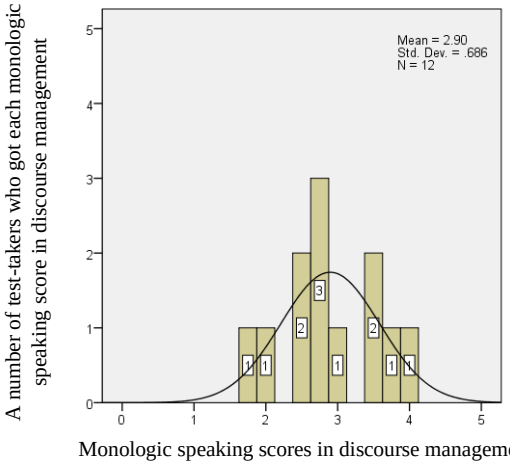


Figure 3.5: Histogram for monologic speaking scores in discourse management – pilot study

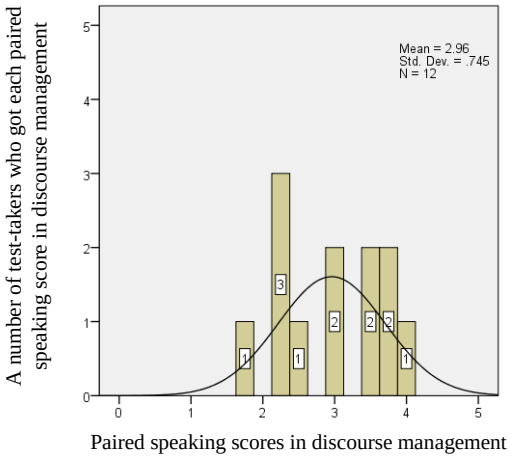


Figure 3.6: Histogram for paired speaking scores in discourse management – pilot study

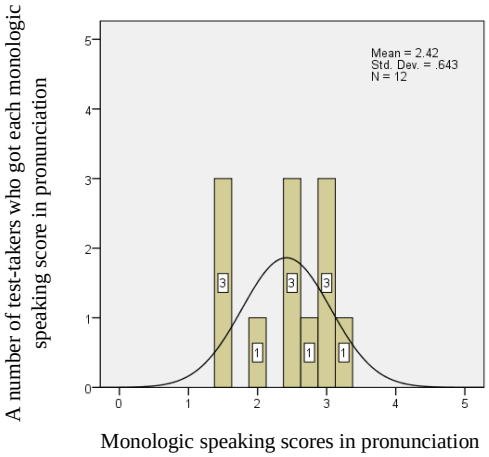


Figure 3.7: Histogram for monologic speaking scores in pronunciation – pilot study

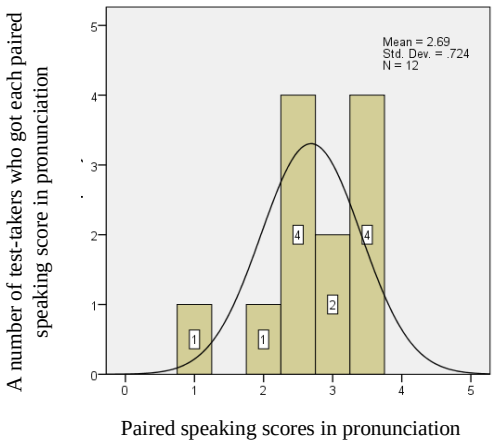


Figure 3.8: Histogram for paired speaking scores in pronunciation – pilot study

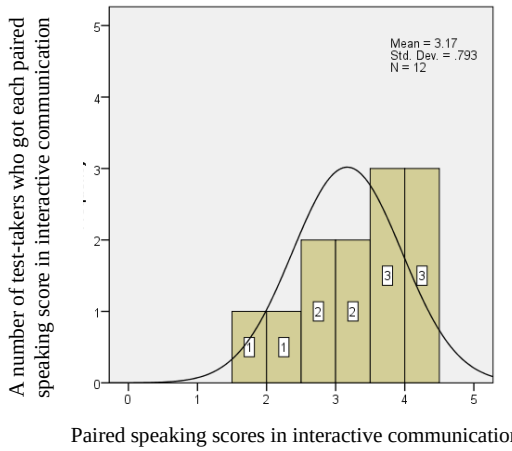


Figure 3.9: Histogram for paired speaking scores in interactive communication – pilot study

As noted above, the sample size of the pilot study was small (N=12), and none of the histograms showed normal distribution. Spearman's rho test was therefore employed to examine the relationship between listening scores and speaking scores in the monologic speaking test and the paired speaking tests, respectively. A summary of the correlations between these variables is presented in Table 3.6.

Table 3.6: Correlations between test-takers' listening scores and analytical scores in the monologic and paired speaking tests (N=12) – pilot study

Speaking Test	Grammar and Vocabulary		Discourse Management		Pronunciation		Interactive Communication	
	Mono	Paired	Mono	Paired	Mono	Paired	Mono	Paired
Spearman's rho	.62*	.43	.40	.71**	.50	.57	-	.77**
Sig. (2-tailed)	.03	.16	.20	.01	.10	.06	-	.00

Note: *. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

As shown in Table 3.6, there was a statistically significant correlation between grammar and vocabulary in monologic speaking scores and listening scores, and between discourse management and interactive communication in paired speaking scores and listening scores. The correlation in the grammar and vocabulary category is not surprising in light of the fact that grammar and vocabulary components usually make up a significant proportion of the total score variance in skills-based tests (e.g., Geranpayeh, 2007; Shiotsu and Weir, 2007). However, of interest are the

statistically significant, positive correlations found between discourse management in paired tests and listening, and between interactive communication in paired tests and listening. These suggest that the better the test-takers' listening was, the more effectively they managed the discussion and interacted in paired tests.

Table 3.7 presents the correlations between test-takers' listening scores and paired speaking scores for each analytical category in shared and non-shared L1 pairs.

Table 3.7: Correlations between test-takers' listening scores and analytical scores in paired speaking tests in shared L1 pairs and non-shared L1 pairs (N=12) – pilot study

	Grammar & Vocabulary		Discourse Management		Pronunciation		Interactive Communication	
	SL	**NSL**	**SL**	**NSL**	**SL**	**NSL**	**SL**	**NSL**
Spearman's rho	.13	.46	.65*	.48	.39	.62*	.52	.58*
Sig. (2-tailed)	.69	.13	.02	.12	.21	.03	.08	.05

Note: *. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).
SL refers to shared L1 pairs.
NSL refers to non-shared L1 pairs.

In the paired speaking tests in shared L1 pairs, the test-takers' speaking scores in the discourse management category showed statistically significant correlations with listening scores at the 0.05 level. In non-shared L1 pairs, the test-takers' speaking score in pronunciation and interactive communication categories showed statistically significant correlations with listening scores. While the first two findings are difficult to interpret at this stage, the third finding could indicate that listening skills are more important in achieving effective communication between non-shared L1 pairs than between shared L1 pairs.

3.3.5.2 Paired speaking test scores awarded to shared and non-shared L1 pairs (RQ2)

The non-parametric Wilcoxon Signed Rank Test was used to investigate the differences between test-takers' paired speaking scores in shared and non-shared L1 pairs in each analytical category. There was no significance at the 0.05 levels,

as shown in Table 3.8.

Table 3.8: Differences between test-takers' analytical scores for paired speaking tests in shared and non-shared L1 pairs (N=12) – pilot study

Category	Mode	Mean	Median	SD	Min	Max	Wilcoxon
Grammar and vocabulary (1–5 points)	**SL**	2.73	2.75	.56	1.50	3.50	Z=-.09 p = .93
	NSL	2.75	2.75	.45	2.00	3.50	
Discourse management (1–5 points)	**SL**	2.94	3.00	.69	1.75	4.00	Z=-.15 p = .88
	NSL	2.90	2.75	.63	2.25	3.75	
Pronunciation (1–5 points)	**SL**	2.63	2.50	.76	1.00	3.50	Z=-.78 p = .44
	NSL	2.75	2.63	.38	2.25	3.50	
Interactive communication (1–5 points)	**SL**	3.13	3.13	.61	1.75	4.00	Z=-.59 p = .56
	NSL	2.92	2.75	.89	1.50	4.25	

Note: SL refers to shared L1 pairs.
NSL refers to non-shared L1 pairs.
The total possible score for each category is 5 points.

However, although there was no statistically significant difference in scores between the two types of pairs, test-takers seemed to achieve slightly better scores on pronunciation when they were paired with non-shared L1 partners, and they achieved slightly better scores on interactive communication when paired with shared L1 partners. This might suggest that they might have made more effort to make pronunciation clearer and produced individual sounds and prosodic features well in order to be understood by their partners when speaking to a non-shared L1 partner. This might support Jenkins' study (2002), which found test-takers from a non-shared L1 background attempted to produce more target-like sounds in the paired interaction than test-takers from a shared L1 background. The interactive communication result might also indicate that shared L1 pairs had more effective communication, possibly because they could understand each other better. These two points will need to be revisited in the main study. If they are verified, then not only do they support the findings of Kachi's (2004) and Fayers and Krasinski's (1987) studies, which concluded that English spoken by shared L1 speakers was more intelligible for a shared L1 listener than English spoken by non-shared L1 speakers, they will have important implications for paired speaking tests.

3.3.5.3 Communication patterns related to interactive listening between shared L1 pairs and non-shared L1 pairs *(RQ3)*

This part presents communication patterns that relate to the ways in which test-takers demonstrated interactive listening through providing comprehension and supportive listening (Ducasse, 2010; Ducasse and Brown, 2009). The former category can be evidenced by candidates' comprehension through (a) their filling a silence by supplying appropriate vocabulary, and (b) demonstrating comprehension. The latter category can be evidenced by (c) candidates' back-channelling (which may be used with gesturing). Some communication patterns related to interactive listening were identified in the pilot study, and they will be presented under two headings: (1) similarities and (2) differences in communication patterns related to interactive listening between shared and non-shared L1 pairs.

3.3.5.3.1 Similarities in communication patterns related to interactive listening between shared and non-shared L1 pairs

Similar communication patterns related to interactive listening between shared and non-shared L1 pairs were identified in terms of *back-channelling* and *support by supplying vocabulary.*

Back-channelling

Candidates in both shared and non-shared L1 pairs used back-channels, for example, *uh huh*, *yeah*, *yes* and *ah*, to display that they were listening to the partner and to encourage their partner to continue to talk. Examples are illustrated in Lines 4 and 6 of Excerpt 1 below.

Excerpt 1

Topic: Friends (T01: Thai female 01; T02: Thai female 02)

L1 **T02:** yeah. mm:: so i think that (0.8) when i- (.) in Thailand
L2 mean i have lots of friends too but i think that (.) like (0.3)
L3 to consider having best friends (0.3) [like close friends
→ L4 **T01:** [uh huh.
L5 **T02:** i don't have a lot [ah::: because i feel that (0.3)
→ L6 **T01:** [uh huh

Support by supplying vocabulary

In Excerpt 2, lines 3 and 6 are examples of showing listening comprehension by filling in a missing word, here "*comfort you*" and "*to have a better feeling? To feel relief*". This feature was suggested as evidence of comprehension in the studies carried out by Ducasse (2010) and Ducasse and Brown (2009). Regarding the stimulated recall interview with T02, she reported that she said those words because she was able to guess what her partner wanted to say but her partner seemed to be having difficulty in searching for the words to express the ideas. Such examples were often obtained from both shared and non-shared L1 pairs.

Excerpt 2

Topic: Friends (T01: Thai female 01, T02: Thai female 02)

L1 **T01:** = best friend (.) come (0.4) ((moving hand to herself)) to L2 stay with you and (0.3) [yes and comfort (0.6) you or::=
→L3 **T02:** [comfort you
L4 **T01:** = or (0.5) ah support you (0.7) to (0.3) to to talk you to to
L5 get a better better ((moving hands)) (0.5) ah: (0.5) bet- ah::
→L6 **T02:** to have a better feeling [to feel relief
L7 **T01:** [yeah yeah

3.3.5.3.2 Differences in communication patterns related to interactive listening between shared and non-shared L1 pairs

The pilot study also identified two differences in communication patterns related to interactive listening between shared and non-shared L1 pairs: *superficial back-channelling* and *causes of listening-related communication breakdown*.

Superficial back-channelling

Although back-channelling was used in both types of pairs as described in Section 3.3.5.3.1, non-shared L1 pairs also seemed to use back-channelling as supportive listening, even when they did not understand their partner, as shown in Excerpt 3. In this excerpt, although T05 seemed to demonstrate interactive listening to I01 by using back-channelling, she looked at the task prompts instead of at her partner's eyes. This observation was supported by the stimulated recall interview with T05. She revealed that she lost her concentration because she did not understand some words used by I01. This was also noted by Rater 1, who awarded I05 and T05 3 and

2 (out of 5) respectively in the interactive communication category. The rater explained her reasoning behind these scores by stating that I01 did not give his partner a turn and that T05 allowed him to dominate the conversation even though she looked bored and confused when he was talking about Formula One.

Excerpt 3

Topic: Professions (I01: Gujarati male 01, T05: Thai female 05)

L1 **I01:** let's start.
L2 **T05:** how dif- difficult is it to be successful in these (0.3)
L3 professions? what do you think which (.) one you choose?
L4 **I01:** (1.2) ah::: i choose er:: i choose from F1 formula 1
L5 [(0.8) formula 1 er::: is difficult to be successful because =
→ L6 **T05:** [aha
L7 **I01:** = ah::: (1.1) are so many competitors (0.4) to (.) be a
L8 successful. er:: i think (0.6) they're twenty or twenty-five
L9 competitors [who interest (2.3) but ah: formula 1is ah::=
→ L10 **T05:** [uh huh

Using superficial back-channelling confirms one of Ducasse and Brown's (2009) findings, that is, that showing supportive listening might not always demonstrate one's listening comprehension to an interlocutor, but a listener might use audible support strategies just to pretend to understand a speaker's message. T05's lack of ability to take an active part in the interaction and the fact that she allowed a partner to dominate the interaction indicated her poor interactive communication skill (Galaczi, 2004).

Causes of listening-related communication breakdown

Communication breakdowns seemed to occur in both shared L1 pairs and non-shared L1 pairs, but the causes for breakdowns seemed to be different. When non-shared pairs had breakdowns, the cause often seemed to be at the decoding stage or at the very beginning of the meaning-building stage of listening. As described in Section 2.4, Field (2008) notes that at the initial decoding stage, listeners have to manage with what they hear that is in the form of acoustic signals and firstly transform it into the sounds of L2, then into vocabulary and phrase in their word and lastly into a conceptual idea. It seems that non-shared L1 pairs often had problems at such an earlier stage of listening. In contrast, shared L1 pairs did not have such fundamental problems in decoding or at the initial stage of meaning

building. When shared L1 pairs had communication breakdowns, the problems seemed to be related to the last two stages of meaning building, i.e., meaning representation and discourse representation (see Section 2.4 for more details), due to a lack of fuller contexts and a fuller explanation.

For example, Excerpt 4 illustrates a communication breakdown between shared L1 speakers, T05 and T06. In line 2, T06 asked clarification questions, "*What kind of friend? What does it mean*?", regarding T05's question in line 1. It is clear from the excerpt that T06 did not have a decoding problem. T06 understood exactly what T05 said, but she had a problem in understanding what T05 meant by the question in that specific context.

Excerpt 4

Topic: Friends (T05: Thai female 05, T06: Thai female 06)

L1 **T05:** ah:::, what kind of friend do you like?
→ L2 **T06:** what kind of friend? what does it mean? huh huh.
L3 **T05:** i mean like ah ((looking at ceiling)) (2.0) friend for travel
L4 ((moving hands to the left)), friend for in your
L5 [work ((moving hands to the right)), ah friend =
L6 **T06:** [work
L7 **T05:** = to go to party with you ((moving hands to the right)).
L8 so something like that

By contrast, when non-shared L1 pairs had a communication breakdown, the breakdown tended to relate to their problem of decoding what the partner said or understanding the partner's utterance as a proposition. As shown in Excerpt 5, A03 said "*excuse me?*" in line 4, which could signal that she might have had a decoding problem rather than a meaning-construction problem. In fact, the stimulated recall interview with A03 confirmed that she did not understand her partner's question at all. However, the question T03 asked was ungrammatical, which must also have contributed to the misunderstanding in this case.

Except 5

Topic: Friends (T03: Thai male 03, A03: Arabic female 03)

L1 **T03:** ok. ah::: (1.0) for the last minute er:: (0.5) which picture do
L2 you think (.) is the (0.9) come is the meaning of your
L3 friend?
→ L4 **A03:** (1.0) excuse me?
L5 **T03:** i mean (.) which picture ((waving a hand around pictures))
L6 you think is meaning meaning (0.3) of a friend?

3.3.6 Summary of the pilot study findings and learning experiences of the pilot study

Although the results of the pilot study are only suggestive due to the small sample size, and they will need to be verified in the main study, some interesting results were indicated.

- Students with higher listening scores tended to get higher paired speaking scores in the discourse management and interactive communication categories.
- Students' listening proficiency seemed to positively correlate with their interactive communication speaking scores in non-shared L1 pairs.
- No statistical difference was observed in paired speaking scores between shared and non-shared L1 pairs.
- Similar communication patterns related to interactive listening between shared and non-shared L1 pairs were identified: use of back-channelling and support by supplying vocabulary.
- Different communication patterns related to interactive listening between shared and non-shared L1 pairs were identified: use of "pretend" back-channelling and causes of misunderstanding.

In light of the experience of the pilot study, it was decided that the research instruments and research procedures for the main study would be revised as follows:

- The test-takers of the pilot study scored very low in the listening test, and the distribution of the listening scores was therefore skewed towards the bottom of the scale. In order to achieve a better distribution, the listening test for the main study would be modified to be at a more appropriate level for the test-takers and to be more meaningful for the purpose of this study. The new listening test for the main study would be devised based on questions derived from Cambridge's Preliminary English Test (PET) and the First Certificate in English (FCE) examinations (see Section 3.4.2.2). It was hoped that this would better distribute students' listening scores.
- The raters in the pilot study were not trained to use the FCE analytic rating scales. As such, although the two raters were experienced raters in other

standardised speaking tests, it was not possible to confirm their understanding of the FCE rating scales. As mentioned in Section 3.3.5.1, the inter-rater reliability of the two raters did not reach the acceptable level suggested by Graham et al. (2012). To enhance the consistency of rating in the main study, it is crucial to train raters to use the FCE analytic rating scales.

- Stimulated recall interviews with raters would be utilised in the main study in order to gain further insights into raters' perception of test-taker performance in shared and non-shared L1 pairs. Hence, raters' stimulated recall interviews with selected video recordings of the test-takers' speaking performance would be carried out in the main study (see Section 3.4.2.2).

3.4 Main study

The main study employed the mixed methods approach illustrated earlier in Figure 3.1. The participants, research instruments, data collection procedures and data analysis of the main study will be presented in this section.

3.4.1 Participants in the main study

The participants in the main study consisted of 40 test-takers and two speaking test raters. Details of the participants are explained below.

3.4.1.1 Test-takers in the main study

Initially, the aim was to gather data from 20 Thai L1 speakers and 20 Arabic L1 speakers from the university's pre-sessional English programme, based on the successful pilot data collection from Thai and Arabic L1 students. However, this plan needed to be revisited for practical reasons. Due to an unexpected decrease in the number of Arabic L1 intake students in the pre-sessional programme just before the time of the main data collection, there were not enough Arabic L1 speaking students for the main study. It was impossible to recruit Arabic L1 students from another university, owing to the complex nature of the research design. For this reason, other L1 backgrounds had to be considered to replace Arabic.

Among others, Urdu L1 speakers, one of the largest groups of the intake students at the university, were thought to be most appropriate, given their comparability of English proficiency with Thai L1 participants. According to Kachru's (2005) three circle model of the spread of English, Pakistan is in the outer circle, while Thailand is in the expanding circle. Although Thailand and Pakistan are in different circles of the spread of English, preliminary interviews carried out while recruiting participants identified that the participating Pakistani students' exposure to English in their home country did not seem to differ from that of the participating Thai students. In selecting Urdu L1 speakers to replace Arabic L1 speakers, it was also considered to be important that Urdu is from a different language family from Thai and that they are phonologically very different. Urdu is a syllable-timed language in which the rhythm is primary to the stress and based upon the long and short syllable arrangement (Shackle, 2001), while Thai is a tonal language in which the pitch of pronouncing each syllable represents a different meaning (Smyth, 2001).

As in the pilot study, participants in the main study were selected by using the purposive sampling method, i.e., targeting Urdu and Thai L1 speakers who had similar IELTS speaking and listening scores.

The study recruited 20 Urdu L1 test-takers and 20 Thai L1 test-takers. Since the literature review suggested some effects of most test-taker characteristics on paired test performance (see Section 2.3.1), the study aimed to control as many test-taker characteristics as possible to avoid potential confounding effects, although none of the test-taker characteristics apart from L1 was the focus of the study. A balanced gender profile in both groups was achieved, including 10 males and 10 females in each language group, and it was decided to make only single-sex pairs. The participants' age was relatively similar. All of them were young adults – most of them were in their 20s (Mean=27.20, SD=2.84). Although the length of time they had been in the UK[2] ranged from one to 18 months (Mean=6.60, SD=4.35),

[2] The main study data was collected at the beginning of the pre-sessional course to minimise potential interference of test-takers' *length of stay in the UK* and *amount of English used inside and outside the classroom*. However, despite the effort to collect the data at the beginning of these students' pre-sessional course, the length of stay in the UK of the 40 students varied from one to 18 months, which is one of the limitations of the study.

most of the test-takers had stayed in the UK for a relatively short period of time (over half of them had stayed for less than 6 months). As for their English ability, all test-takers' IELTS overall scores ranged from Band 4.5 to Band 5.5 (Mean=5.33, SD=.27), with speaking scores from Band 5.0 to Band 6.0 (Mean=5.61, SD=.35), and listening scores from Band 5.0 to Band 6.5 (Mean=5.28, SD=.39). The two L1 groups of test-takers seemed to have comparable overall, speaking and listening proficiency in IELTS (see Section 11.2.1 in Appendix 11 for more details). The paring methods used in the main study and paired test-takers' characteristics are presented in Section 3.4.3.

Like the pilot study, a demographic questionnaire in the main study also asked about the test-takers' perceived familiarity with the English spoken by shared and non-shared L1 speakers. As detailed in Appendix 11, Urdu L1 test-takers reported that they were more familiar with English spoken by Urdu L1 speakers (Mean=4.2 of the familiarity scale of 1-5) than Thai L1 speakers (Mean=3.6). Similarly, Thai L1 test-takers indicated that they were more familiar with English spoken by Thai L1 speakers (Mean=3.85) than Urdu L1 speakers (Mean=2.25), but their ratings were in general lower than the Urdu L1 speakers' ratings across both categories.

3.4.1.2 Speaking test raters and rater training

The same two raters, both native speakers of English, who participated in the pilot study took part in the main study. As described earlier, in Section 3.3.1.2, they are experienced speaking raters, and are not familiar with either Urdu or Thai L1s (see Section 3.3.1.2 for more background information of the raters).

As suggested in Section 3.3.6, a face-to-face training session for the use of the FCE analytic rating scales was provided to the raters in order to enhance the reliability of their rating. The latest publicly available version of the Cambridge FCE speaking criteria (UCLES, 2012), which has four analytical categories: *grammar and vocabulary*, *discourse management*, *pronunciation* and *interactive communication* (for more details about the four categories, see Section 3.3.2.5 and Appendix 7) was used in the main study. The latest version has the same descriptors and the same six levels (from 0 to 5) as the 2008 version, but it does not allow raters to award a half mark (0.5). The main study therefore also allowed only full marks.

In the rater training session, the rating scale was firstly explained to the raters in order to establish a mutual frame of understanding for awarding speaking scores. Secondly, the raters watched a video recording of the monologic speaking test performance, which was taken from the pilot study. The raters were then asked to award the test-taker's analytical scores and provide reasons for their judgements. The awarded scores were compared with the scores gained from the researcher's analysis of the test performance against the FCE rating descriptors. It should be noted that the training session was not official training. It was similar to a discussion among peers and was conducted to raise the raters' awareness of various components of the FCE rating scales, since both raters were experienced speaking test examiners in other standardised tests although they were unfamiliar with the FCE rating scales. The reasons for awarding scores were discussed until a mutual understanding was arrived at between the researcher and the two raters. Next, the raters watched another video recording of the monologic speaking task and awarded scores as a practice exercise. The scores from the raters were compared and discussed again. The rater training for the paired speaking task was operated in a similar way to that for the monologic speaking task by using two video recordings of shared L1 pairs and two video recordings of non-shared L1 pairs from the pilot study. The rater training lasted approximately an hour.

3.4.2 Research instruments for the main study

The following section explains the research instruments which were used in the main study. Since the pilot study demonstrated the successful use of the background questionnaire (see Section 3.3.2.1 and Appendix 1), the speaking tests (see Section 3.3.2.3 and Appendices 3 and 4) and the stimulated recall interview with test-takers (see Section 3.3.2.4), these research instruments were employed in the main study without any modifications. As suggested in Section 3.3.6, the listening test used in the pilot study needed modifying, and it was decided that a stimulated recall interview with speaking raters would be included. The following sections detail the new listening test and the raters' stimulated recall interview used in the main study (for the rest of the instruments, see Section 3.3.2).

3.4.2.1 Listening test

As discussed in Section 3.3.6, since the distribution of the listening scores of the pilot study was skewed towards the bottom of the scale, it was decided to make the test easier. The new listening test for the main study was devised based on items taken from Cambridge's Preliminary English Test (PET) and the First Certificate in English (FCE) examinations. The PET is a B1 test, whose listening section aims to assess test-takers' ability to follow and comprehend a range of spoken materials as well as announcements and discussions related to everyday life (UCLES, 2014, p. 3). As noted in Section 3.3.2.2, the FCE is a B2 test, and the listening section focuses on assessing test-takers' ability to understand the meaning of a range of spoken language material together with lectures, radio broadcasts and speeches as talks (UCLES, 2015, p. 3). It was hoped that combining the listening test items from both tests would better distribute students' listening scores.

The listening test (see Appendix 8) for the main study consisted of 39-test items, each item being worth one mark. Nineteen-test items were taken from PET practice material published by the University of Cambridge ESOL Examinations (Cambridge ESOL, 2008). These consisted of multiple-choice (13-test items) and yes-no questions (6-test items). Twenty-test items were derived from the FCE practice material published by the University of Cambridge ESOL Examinations (Cambridge ESOL, 2009). These consisted of multiple-choice (15-test items) and multiple-matching (5-test items) tasks.

Since the new listening test included only selected items from the PET and the FCE, the reliability of the listening test needed to be established first to ensure that the resulting scores were consistent and dependable (Brown, 2004, p. 20). Prior to the main study, the new listening test was trialled with 30 students who were in a pre-sessional English language programme at the University of Bedfordshire. The trial was administered to a different population of the students from those participating in the main study. After they finished the listening test, all listening test answer sheets were marked by the researcher.

Cronbach's alpha was utilised to investigate the reliability of the listening test. The reliability coefficient of the 39 items was .90 (see Appendix 9 for the full details). According to McNamara (2000, p. 62), a reliability coefficient of 0.90 or

better is what researchers normally look for with a comprehension test. Although the alpha value seemed acceptable, the discrimination levels of some items were inadequate. The item discrimination is "the extent to which an item differentiates between high- and low-ability test-takers" (Brown, 2005, p. 59). Values of item discrimination (discrimination index) level of .20 or higher (r≥ .20) are acceptable to measurement. An item with an item discrimination value of less than .20 means that it is not possible to discriminate between strong and weak test-takers, and it is suggested that such items should be deleted from a listening test. An item with a negative discrimination value is seriously problematic, and it is strongly recommended that such items should be discarded.

Two-test items which had negative item discrimination values were discarded from the listening test. These were Items 1 and 3. Table 3.9 shows their values of item facility and item discrimination. It seems that both items had very high facility values which might have caused the low item discrimination indices.

Table 3.9: Values of item facility and item discrimination of discarded listening test items – main study

Item	Item facility (p)	Item discrimination (r)	Reason for discarding the items
1	.93	-.05	Very easy and unable to discriminate between test-takers with a high listening proficiency level and test-takers with a low listening proficiency level
3	.87	-.09	

After deleting the two-test items (1 and 3) from the listening test, the reliability of the listening test (37-test items) increased to .91 (see Appendix 9).

3.4.2.2 Stimulated recall interviews with raters

In the main study, stimulated recall interviews with raters were added in the hope of gaining greater understanding of what raters noticed and attended to while rating. This was carried out after all test scores were gathered by the two raters.

Eight video-recorded performances were selected as stimuli and the raters were individually interviewed while they watched each of the eight videos of paired tests once again. As shown in Table 3.10, the eight video recordings included four shared L1 pairs and four non-shared L1 pairs, consisting of four Urdu L1 speakers and four Thai L1 speakers. Their gender profiles were also balanced. The selection

was also informed by the researcher's preliminary analysis during the test administration and test-takers' stimulated recall interviews, and it was hoped that these pairs would cover some typical interactional features observed in the 40 participants. The features preliminary identified at that stage included supplying related vocabulary, back-channelling, demonstrating comprehension and dealing with communication problems.

Table 3.10: Video recordings selected for the raters' stimulated recall interviews

Test-taker	Gender	English proficiency based on IELTS examination score		Type of pair	
		Speaking	Listening	Shared L1 pair	Non-shared L1 pair
P03	Male	6.0	6.0	P03P04	P03T03
P04	Male	6.0	6.0		
T03	Male	5.5	6.0	T03T04	P04T04
T04	Male	6.0	5.5		
P19	Female	6.0	5.0	P19P20	P19T19
P20	Female	6.0	5.0		
T19	Female	5.5	5.0	T19T20	P20T20
T20	Female	6.0	5.0		

While watching the eight selected video recordings once again individually with the researcher, the raters were asked to report the reasons for awarding scores in each of the four analytical categories and any salient interactional features they noticed. They were given a rating sheet that included their written comments with which they had originally rated the selected test-takers' performance. While watching a video recording, the raters were allowed to stop it whenever they wanted to speak about their reasons for awarding a score in each analytical category and to point out any interactional features which they thought worth commenting on. In addition, the researcher paused the video recording when she wanted to know about the raters' perception of any particular test-taker performance, and asked questions such as "*Why do you think test-taker X did (or said) this?*" All the interviews were audio recorded, transcribed and coded according to the themes that emerged. The results were then used to triangulate and elaborate on the findings from the CA analysis of test-taker performance. The analysis method used to code the retrospective interviews will be detailed in Section 3.4.4.

3.4.3 Data collection procedures for the main study

The data collection procedures for the main study followed almost the same procedures used in the pilot study (see Section 3.3.3), with a few modifications. The researcher gained access to students' background information (e.g., L1, country of origin and standardised English test records) via staff members at the Language Centre at the University of Bedfordshire. As noted in Section 3.4.1.1, the researcher decided to select Urdu L1 students instead of Arabic L1 students in the main study due to an unexpected decrease in the number of new Arabic L1 students. The researcher then selected participants by using the purposive sampling method, which aimed to recruit 20 Thai (10 males and 10 females) and 20 Urdu (10 males and 10 females) L1 speakers with similar levels of English speaking and listening proficiency according to their recent IELTS scores.

Before starting the data collection, test-takers' consent to participate in the study was sought and they all signed a consent form. After that, they were asked to complete a background questionnaire (see Section 3.3.2.1 and Appendix 1). The test-takers' information gathered from the questionnaires was used to attempt to control for additional test-taker characteristics (i.e., age, gender and English speaking and listening proficiency) when pairing test-takers for the paired speaking tasks to control for possible confounding variables.

The researcher gave the test-takers ID numbers by using *P* for Pakistani and *T* for Thai and using a number after *P* and *T* to identify each test-taker, e.g., *P01*, *T01*, *P02*, *T02*, and so on, up to *P20* and *T20*. Test-takers were divided into ten groups based on their background information taken from the questionnaire. Each group consisted of four test-takers, who had the same gender, a similar age range and similar English speaking and listening proficiency levels, of which two were Thai L1 speakers and two were Urdu L1 speakers.

It should be noted that because there was a larger number of test-takers (40 test-takers) in the main study, the listening test and the monologic speaking test were administered on the same day and the two paired speaking tests were administered over a period of two and a half days (for the data collection procedures in listening and monologic speaking tests, the main study repeated the pilot data collection procedures presented in Section 3.3.3).

On each of the first two days of data collection of paired speaking performance, 16 test-takers participated: two groups (4 shared L1 and 4 non-shared L1 pairs) in the morning and the next two groups (4 shared L1 and 4 non-shared L1 pairs) in the afternoon. On the third day, the remaining two groups (4 shared L1 and 4 non-shared L1 pairs) were tested.

The details of the pairing and grouping of the test-takers as well as paired speaking test tasks are illustrated in Table 3.11.

Table 3.11: Pairing and grouping the test-takers and paired speaking test tasks used – main study

No.	Code	L1	Country of origin	Gender	Age	IELTS score		Type of pair		Group
						Speaking	Listening	Shared L1	Non-shared L1	
1	P01	Urdu	Pakistan	Male	25	6.0	6.0	P01P02	P01T01	1
2	P02	Urdu	Pakistan	Male	25	5.5	5.0			
3	T01	Thai	Thailand	Male	27	6.0	5.0	T01T02	P02T02	
4	T02	Thai	Thailand	Male	29	5.5	5.5			
5	P03	Urdu	Pakistan	Male	28	6.0	6.0	P03P04	P03T03	2
6	P04	Urdu	Pakistan	Male	28	6.0	6.0			
7	T03	Thai	Thailand	Male	27	5.5	5.0	T03T04	P04T04	
8	T04	Thai	Thailand	Male	25	6.0	5.0			
9	P05	Urdu	Pakistan	Male	24	5.0	6.5	P05P06	P05T05	3
10	P06	Urdu	Pakistan	Male	24	5.0	6.0			
11	T05	Thai	Thailand	Male	25	5.5	5.0	T05T06	P06T06	
12	T06	Thai	Thailand	Male	27	5.0	5.0			
13	P07	Urdu	Pakistan	Male	28	5.5	5.0	P07P08	P07T07	4
14	P08	Urdu	Pakistan	Male	33	6.0	5.0			
15	T07	Thai	Thailand	Male	28	5.5	5.0	T07T08	P08T08	
16	T08	Thai	Thailand	Male	33	5.5	5.5			
17	P09	Urdu	Pakistan	Male	27	5.5	5.5	P09P10	T09T10	5
18	P10	Urdu	Pakistan	Male	31	6.0	5.0			
19	T09	Thai	Thailand	Male	28	5.5	5.5	P09T09	P10T10	
20	T10	Thai	Thailand	Male	28	6.0	4.5			
21	P11	Urdu	Pakistan	Female	29	5.5	5.0	P11P12	P11T11	6
22	P12	Urdu	Pakistan	Female	25	6.0	5.5			
23	T11	Thai	Thailand	Female	24	5.5	5.0	T11T12	P12T12	
24	T12	Thai	Thailand	Female	24	6.0	5.0			
25	P13	Urdu	Pakistan	Female	28	5.5	5.0	P13P14	P13T13	7
26	P14	Urdu	Pakistan	Female	30	5.5	5.5			
27	T13	Thai	Thailand	Female	26	5.0	5.5	T13T14	P14T14	
28	T14	Thai	Thailand	Female	25	5.0	5.0			
29	P15	Urdu	Pakistan	Female	23	5.5	5.5	P15P16	P15T15	8
30	P16	Urdu	Pakistan	Female	27	5.5	5.0			
31	T15	Thai	Thailand	Female	24	5.0	5.0	T15T16	P16T16	
32	T16	Thai	Thailand	Female	23	5.0	5.0			
33	P17	Urdu	Pakistan	Female	24	6.0	5.5	P17P18	P17T17	9
34	P18	Urdu	Pakistan	Female	28	5.5	5.0			
35	T17	Thai	Thailand	Female	26	6.0	5.0	T17T18	P18T18	
36	T18	Thai	Thailand	Female	29	5.5	5.0			
37	P19	Urdu	Pakistan	Female	28	6.0	5.0	P19P20	P19T19	10
38	P20	Urdu	Pakistan	Female	33	6.0	5.0			
39	T19	Thai	Thailand	Female	34	5.5	5.0	T19T20	P20T20	
40	T20	Thai	Thailand	Female	28	6.0	5.0			

Note: P refers to Pakistani and T refers to Thai test-takers.

As shown in Table 3.11, the researchers attempted to pair participants with those with similar IELTS speaking and listening scores to prevent their speaking and listening proficiency from confounding the study findings. It was, however, not always possible to match students with exactly the same speaking and listening band scores. In such cases, the priority went to their speaking scores (not different by more than 0.5 band score) and then listening scores (not different by more than 1.0 band score). For example, P01, who was awarded IELTS speaking and listening Band 6.0, was paired with P02, who was awarded Bands 5.5 and 5.0 respectively. P01 was also paired with T01, who had gained speaking and listening bands 6.0 and 5.0. P02 was paired with T02, who had speaking and listening proficiency at the level of Band 5.5. Although all possible efforts were made, there was one pair (P05T05) who had to have a difference of 1.5 bands in relation to their listening scores.

The test-takers in the odd-numbered groups (Group1 and Group 3, and so on up to Group 9) did task A with a shared L1 partner and did task B with a non-shared L1 partner. The test-takers in the even-numbered groups (Group 2 and Group 4, and so on up to Group 10) did task B with a shared L1 partner and did task A with non-shared L1 partner. The order of shared and non-shared L1 pairs was also counterbalanced in order to prevent a possible practice effect. As in the pilot study, all test-takers had stimulated recall interviews immediately after completing paired tasks. Such procedures in the main study were the same as in the pilot study, as explained in Section 3.3.3.

All speaking recordings of the test-takers' speaking test performance (40 monologic and 40 paired recordings) had been copied onto DVDs, and these were sent to the two raters by post. Like in the pilot study, the order of recordings in DVDs was carefully arranged to prevent the previous performance of the same candidate from influencing raters' judgement of the same candidate's performance in another task. More details about the order of all video recordings in DVDs can be seen in Table 3.12.

Table 3.12: Order of video recordings of the test-takers' speaking performance on DVDs

Order of video recording on a DVD	Monologic speaking performance		Paired speaking performance				Group
	DVD 1		DVD 2		DVD 3		
	Rater 1	Rater 2	Rater 1 (Task A)	Rater 2 (Task B)	Rater 1 (Task B)	Rater 2 (Task A)	
1	P01	T01	P01P02	P01T01	P02T02	T01T02	1
2	P02	T02	T01T02	P02T02	P01T01	P01P02	
3	P03	T03	P03T03	P03P04	T03T04	P04T04	2
4	P04	T04	P04T04	T03T04	P03P04	P03T03	
5	P05	T05	P05P06	P05T05	P06T06	T05T06	3
6	P06	T06	T05T06	P06T06	P05T05	P05P06	
7	P07	T07	P07T07	P07P08	T07T08	P08T08	4
8	P08	T08	P08T08	T07T08	P07P08	P07T07	
9	P09	T09	P09P10	P09T09	P10T10	T09T10	5
10	P10	T10	T09T10	P10T10	P09T09	P09P10	
11	P11	T11	P11T11	P11P12	T11T12	P12T12	6
12	P12	T12	P12T12	T11T12	P11P12	P11T11	
13	P13	T13	P13P14	P13T13	P14T14	T13T14	7
14	P14	T14	T13T14	P14T14	P13T13	P13P14	
15	P15	T15	P15T15	P15P16	T15T16	P16T16	8
16	P16	T16	P16T16	T15T16	P15P16	P15T15	
17	P17	T17	P17P18	P17T17	P18T18	T17T18	9
18	P18	T18	T17T18	P18T18	P17T17	P17P18	
19	P19	T19	P19T19	P19P20	T19T20	P20T20	10
20	P20	T20	P20T20	T19T20	P19P20	P19T19	
21	T01	P01					
22	T02	P02					
23	T03	P03					
24	T04	P04					
25	T05	P05					
26	T06	P06					
27	T07	P07					
28	T08	P08					
29	T09	P09					
30	T10	P10					
31	T11	P11					
32	T12	P12					
33	T13	P13					
34	T14	P14					
35	T15	P15					
36	T16	P16					
37	T17	P17					
38	T18	P18					
39	T19	P19					
40	T20	P20					

DVD 1 showed the video recordings of test-takers' monologic speaking performance. Rater 1 received a DVD of test-takers' monologic speaking

performance. The order of the test-takers was as follows: the Urdu L1 test-takers were first ([1]P01, [2]P02, [3]P03 and so on up to [18]P18, [19]P19 and [20]P20), followed by the Thai L1 test-takers ([21]T01, [22]T02, [23]T03 and so on up to [38]T18, [39]T19 and [40]T20). Rater 2 received a DVD which began with the performances of Thai L1 test-takers ([1]T01, [2]T02, [3]T03 and so on up to [18]T18, [19]T19 and [20]T20), followed by Urdu L1 test-takers ([21]P01, [22]P02, [23]P03 and so on up to [38]P18, [39]P19 and [40]P20).

DVD 2 given to Rater 1 included paired performances in Task A, while DVD 2 for Rater 2 included paired performances in Task B. In the DVD of Task A, odd-numbered groups were the shared L1 pairs (Group 1: [1]P01P02, [2]T01T02; Group 3: [5]P05P06, [6]T05T06; and so on up to Group 9: [17]P17P18, [18]T17T18) and even-numbered groups were the non-shared L1 pairs (Group 2: [3]P03T03, [4]P04T04; Group 4: [7]P05T05, [8]P06T06; and so on up to Group 10 [19]P19T19, [20]P20T20). In the DVD of Task B , odd-numbered groups were the non-shared L1 pairs (Group1: [1]P01T01, [2]P02T02; Group 3: [5]P05T05, [6]P06T06; and so on up to Group 9: [17]P17T17, [18]P18T18) and even-numbered groups were the shared L1 pairs (Group 2: [3]P03P04, [4]T03T04; Group 4: [7]P07P08, [8]T07T08; and so on up to Group 10: [19]P19P20, [20]T19T20).

DVD 3 given to Rater 1 contained paired performances in Task B, while DVD 3 for Rater 2 contained paired performances in Task A. As shown in Table 3.11, the DVD showing performances in Task B began with the second performance, followed by the first performance of each group; the odd-numbered groups were the non-shared L1 pairs and the even-numbered groups were the shared L1 pairs. DVD 3 for Rater 2 began with the second performance, followed by the first performance of each group; the odd-numbered groups were the shared L1 pairs and the even-numbered groups were the non-shared L1 pairs. As such, an effort was made to prevent raters from being influenced by the same test-takers' performance in another task, as well as cancelling out possible order effects.

After the raters had finished awarding speaking scores to test-takers, they were invited to take part in a stimulated recall interview individually. The raters were asked to watch each of the eight selected videos and to report on their reasons for awarding scores and on their perceptions of the paired performance. All interviews were audio recorded and transcribed for analysis. The selection of the eight video recordings and the way in which the interviews were carried out are described earlier, in Section 3.4.2.2, and the analysis method will be illustrated in Section 3.4.4.

3.4.4 Data analysis of the main study

Test scores

Quantitative data (demographic data from the questionnaire, listening and speaking scores) were analysed using the SPSS program. Spearman correlation was used to examine the strength of the correlations between the listening test scores and the analytical scores of monologic and paired speaking tests in the whole group, in shared and non-shared L1 pairs separately (to answer ***RQ1***). The non-parametric Wilcoxon Signed Rank Test was used to examine the difference between the analytical scores of paired speaking tests in shared and non-shared L1 pairs (to answer ***RQ2***). To ensure speaking raters' scoring validity, inter-rater reliability[3] and inter-rater absolute agreement were checked (Graham et al., 2012) in the main study. The reliability results will be reported in the next chapter in Section 4.2.

Test-takers' spoken performance in paired tests and stimulated recall interviews

All video recordings of paired speaking performance were transcribed following CA conventions (Atkinson and Heritage, 1984). CA was then carried out to explore communication patterns in the paired tests which were related to test-takers' listening abilities and their L1s (to address ***RQ3***). The steps taken for the data analysis were the same as those taken in the pilot study. Following the CA conventions, the analysis done here also regarded repeated listening to production

[3] The inter-rater reliability refers here to "the measurement of the consistency between evaluators in the *ordering* or *relative standing* of performance ratings, regardless of the absolute value of each evaluator's rating" (Graham et al., 2012, p. 5).

of a transcript as an important part of discovery through analysis (Hutchby and Wooffitt, 1998, p. 73; Levinson, 1983, pp. 286-294). Although the researcher had some potential analysis categories in mind (see below), the analytic stance of "*unmotivated looking*" was maintained to some extent to avoid a description of the given interaction being biased by the literature, and this helped to uncover recurrent and systematic features of group interactions to elaborate the results of the quantitative analyses.

Salient conversational features therefore emerged from the data as the analysis proceeded. Although it is not always recommended[4] (Schegloff, 1993), the transcripts were then coded according to the salient interactional features that emerged. Some quantification has indeed been attempted in relation to these types of communicative events, and inter-coder reliability related to such quantification is provided wherever appropriate. It should be noted that the quantification was only utilised to support the CA findings. As Schegloff (1993) notes, CA is not fully compatible with quantification, since evidence from a single case could have a significant meaning in CA studies. Quantification is nevertheless useful to provide a more generalisable picture of the interactional features under investigation; however, it is crucial to use this in a meaningful way. It is suggested that quantification should be used together with single case analysis in order to understand the "environment of relevant possible occurrence" (Schegloff, 1993, p. 106).

As suggested by Ducasse (2010), Ducasse and Brown (2009) and May (2007), whose work is reviewed in Chapter 2 (Section 2.4.1), a listener's interactive listening can be measured to some extent from a listener's behaviour or response. Hence, this study analysed the test-takers' interactive listening through their interactional behaviour or response in the paired speaking test tasks. In order to support the researcher's interpretation of these communicative events, comments

[4] Coding and quantification in CA can be controversial (Lazaraton, 2002, pp. 82-87). Schegloff (1993) argues that the quantification of conversational data is premature if our understanding of the target features that we wish to count and the environments in which they occur are still incomplete. However, even if "quantification is not a substitute for analysis" (Schegloff, 1993, p. 114), quantification plays a potentially useful role in offering a more generalisable picture of the data by enabling statistical investigation into the occurrence of the properties in question. This study therefore applied coding on CA transcripts.

obtained from test-takers' stimulated recall interviews were used.

The interpretation and the analysis of the interactional transcripts were to some extent based on Ducasse's (2010) findings on interactive listening together with Gass and Varonis's (1991) findings on communication patterns and communication problems related to listening proficiency. However, pre-formulated coding is not compatible with CA (Lazaraton, 2002) and therefore no coding categories were formulated before starting to analyse the data in the main study.

After the researcher coded the entire data according to the themes that emerged, the inter-coder reliability was examined by another coder who coded the entire data set. The researcher trained a PhD student in language testing in how to identify the themes that emerged. The transcriptions of the test-takers' performance in pairs were put into a Word file, and their video recordings were given to the other coder. She was requested to watch the test-takers' video recordings while reading the transcriptions. She was then asked to copy the transcription segments which fitted with each coding scheme to an Excel file. The coders were given the coding scheme developed by the researcher, and she individually coded the data set. To examine the accuracy of the coding, the inter-coder agreement was calculated from the number of times the coders agreed on coding divided by the total number of codes (Graham et al., 2012). The inter-coder reliability figures are provided where appropriate in Chapter 5. Any disagreements between the two coders were discussed until agreement was reached. For example, there was disagreement about the category of misunderstanding because of having a different cultural background (see Section 5.1.2.3, Excerpt 22), which one coder thought should be coded in this category, while the other coder did not recognise this misunderstanding between both test-takers. Data gained from a stimulated recall interview with one of the test-takers in this pair disclosed that his partner did not understand what he intended to deliver because of their different cultural background. Therefore, the coders agreed to code it as coming within the category of misunderstanding because of having a different cultural background.

After confirming the accepted results of coding, the researcher identified (1) the communication patterns that occurred both in the shared and in the non-shared L1 pairs, (2) the communication patterns that occurred either in the shared L1 pairs

or in the non-shared L1 pairs, and (3) communication problems observed in both pairs. Therefore, these communication patterns were classified under three main themes. The resulting categories are summarised in Table 3.13 below. For more details and excerpts, see Chapter 5, Section 5.1.

Table 3.13: Salient interactional features that emerged from CA analysis

Main theme	Sub-theme
1: Similarities in communication patterns between shared and non-shared L1 pairs	**1.1: Supplying relevant vocabulary** A test-taker provides a relevant word or phrase while listening to a partner.
	1.2: Demonstrating comprehension A test-taker responds to a partner's message with a relevant contribution.
	1.3: Back-channelling A listener test-taker produces a speech sound in order to provide supporting feedback to a speaker.
2: Differences in communication patterns related to interactive listening between shared and non-shared L1 pairs	**2.1: Causing communication problems** **2.1.1 Non-engagement** – A test-taker does not participate in a conversation. **2.1.2 Miscommunication** – A test-taker provides an irrelevant response or partly relevant response, leading to a clarification request or confirmation checks.
	2.2: Understanding unclear utterances and incorrect word use A test-taker can understand their partner even when he/she produces an unclear utterance, incorrect words and wrong grammar.
	2.3: Misunderstanding because of different cultural background A test-taker misunderstands his/her partner because they do not share the same cultural background (e.g. religious constraints).
3: Additional interactional features between test-takers in shared and non-shared L1 pairs	**3.1: Providing L1 back-channelling while listening**
	3.2: Inserting an L1 word while speaking

The first category was the similarities in communication patterns between the shared L1 pairs and the non-shared L1 pairs. As stated earlier, in Section 2.4.1, Ducasse (2010) divides interactive listening, which contributes to successful interaction, into two subcategories: comprehension and supportive listening. Comprehension can be evidenced by test-takers' (1) supplying appropriate vocabulary and (2) demonstrating comprehension by responding to a partner's message with a relevant contribution. Supportive listening can be evidenced by (3) candidates' back-channelling (which may be used with gesturing and does not always indicate the listeners' understanding). These three types of performance

theme also emerged in this study as salient features, and they seemed to be utilised in both shared and non-shared L1 pairs.

The second category was the differences in communication patterns between the two types of pair, whose sub-themes included three subcategories: (1) causing communication problems (non-engagement and miscommunication), (2) understanding unclear utterances and incorrect word use, and (3) misunderstanding because of different cultural background.

The subcategory of causing communication problems was identified, and this was then mapped onto Gass and Varonis's (1991) communication problem classification: non-engagement and miscommunication (see definition of these types in the Glossary Section). The two types of causing communication problems identified in this study were also thought to be associated with test-takers' listening proficiency. Non-engagement was observed by test-takers who were quiet or provided only back-channelling as a response to a Wh-question and then shifted to a new topic. Miscommunication was observed by test-takers who provided an irrelevant response, a partly relevant response, a clarification request (e.g., *What do you mean?, What?, Pardon?, Sorry?*) and a confirmation check (*e.g., You mean*…).

Test-takers' stimulated recall interview data were transcribed. The transcripts were then organised in conjunction with the above CA transcripts to check whether the interview data supported the CA analysis. Since the number of the test-takers' comments was limited, and the purpose of the data was just to triangulate the researcher's CA interpretation, no special coding was carried out for the test-taker interview data.

Raters' stimulated recall interviews

Raters' stimulated recall interviews were thematically analysed. The transcripts were put into a Microsoft Excel spreadsheet and, using the sorting function of the Excel program, the data were coded. The small sample size selected for this part of analysis allowed the researcher to use Excel for this purpose, rather than using qualitative data analysis software such as NVivo.

The transcripts were coded so that whenever the raters mentioned their perception of shared and non-shared L1 pairs' interaction related to their interactive

listening, their comments were highlighted and categorised. All the data were transcribed and analysed by the researcher. No inter-coder reliability was checked for this analysis, as this analysis was supplementary to the above CA analysis in that it elaborated on the CA results by offering insights from raters. The themes that emerged are summarised in Table 3.14 below.

Table 3.14: Themes that emerged from raters' retrospective interview data

Main theme	Sub-theme
1. Shared L1 pairs seemed more relaxed and more interactive than non-shared L1 pairs	-
2. Shared L1 pairs' increased mutual understanding and their attempts to solve communication problems	-
3. Additional features for successful interactive communication	3.1: Pronunciation
	3.2: Eye-contact
	3.3: Use of fillers
	3.4: Interactive listening
	3.5: Intelligible word use
	3.6: Confidence

CHAPTER 4 Results of Test Score Analysis

This chapter will present the results of the test score analysis done in the main study. Results from the following data analyses will be presented:

- descriptive statistics for listening and speaking scores;
- inter-rater reliability in awarding speaking scores;
- correlations (a) between listening and monologic speaking scores, (b) between listening and paired speaking scores in the whole group and (c) in shared and non-shared L1 pairs separately (to address RQ1); and
- differences in paired speaking scores when test-takers were paired with shared L1 partners as compared to when they are paired with non-shared L1 partners (to answer RQ2).

Lastly, a summary of the main findings of quantitative analysis will be provided.

4.1 Descriptive statistics for listening and speaking scores

This section presents the descriptive statistics, including the histograms, for listening and speaking scores. The score distributions of both tests were used to make a decision about whether to use a parametric or a non-parametric test to examine the research questions of the main study.

The descriptive statistics for listening and speaking scores are shown in Table 4.1 and histograms are shown in Figures 4.1–4.8. It should be pointed out that the monologic and paired speaking scores were calculated from the average scores of the two raters (see the rationale for this in Section 4.2). As there were two tasks for the paired test, the paired test scores were also computed from the average scores of the two tasks.

Table 4.1: Descriptive statistics for listening scores and analytical scores of monologic and paired speaking tests

Test	Category		Min	Max	Mean	Median	SD
Listening	-		15.00	30.00	21.78	21.50	3.56
Speaking	Grammar and vocabulary (1–5 points)	Mono	.50	5.00	3.58	4.00	1.48
		Pair	.75	5.00	3.36	3.38	1.09
	Discourse management (1–5 points)	Mono	.50	5.00	3.33	4.00	1.43
		Pair	1.00	4.75	3.25	3.50	1.05
	Pronunciation (1–5 points)	Mono	0.00	5.00	3.28	3.50	1.39
		Pair	1.25	5.00	3.18	3.13	1.02
	Interactive communication (1–5 points)	Mono	-	-	-	-	-
		Pair	.75	5.00	3.19	3.50	1.06

Note: The total possible score for the listening test is 37.
The total possible score for the monologic speaking test is 15.
The total possible score for the paired speaking tests is 20.
Mono refers to monologic speaking test.
Pair refers to paired speaking tests.

In the 37-item listening test, the minimum and maximum scores were 15 and 30, with a mean score of 21.78 (SD=3.56). For the monologic speaking test, the scores in grammar and vocabulary, discourse management and pronunciation categories ranged from 0.00 to 5.00, with the average scores being 3.58, 3.33 and 3.28, respectively.

For the paired speaking test, the scores in grammar and vocabulary, discourse management, pronunciation and interactive communication categories ranged from 0.75 to 5.00. The mean scores for all four categories were 3.36, 3.25, 3.18 and 3.19, respectively.

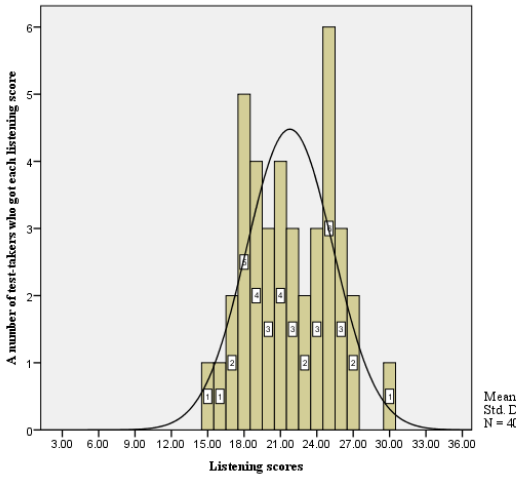


Figure 4.1: Histogram for test-takers' listening scores

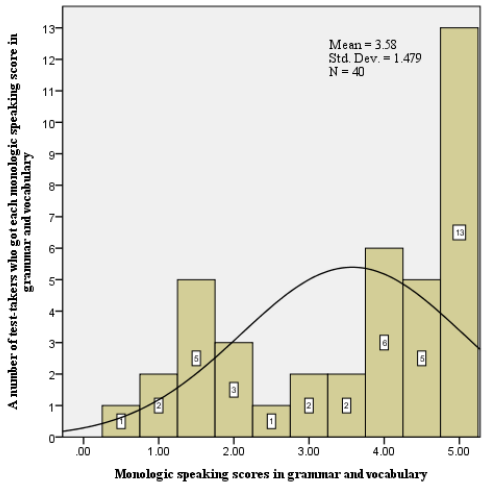


Figure 4.2: Histogram for test-takers' monologic speaking scores in grammar and vocabulary

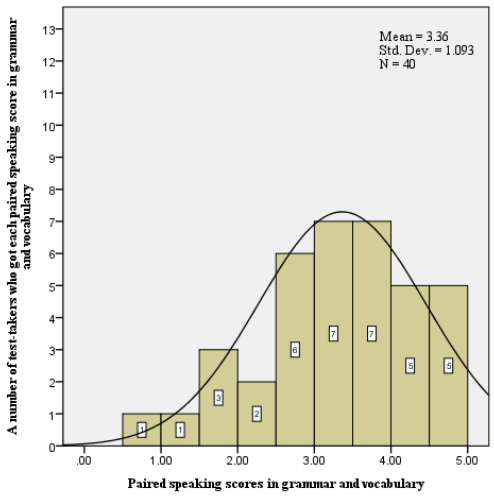


Figure 4.3: Histogram for test-takers' paired speaking scores in grammar and vocabulary

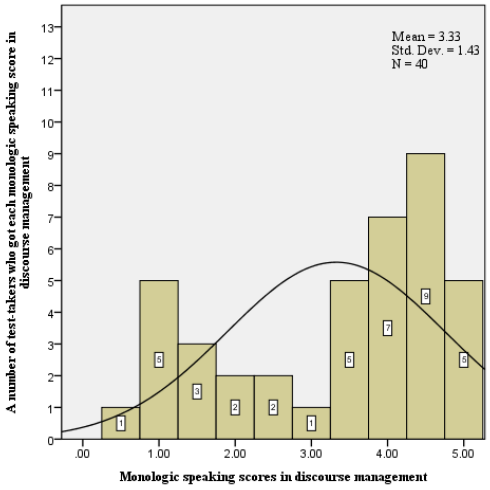


Figure 4.4: Histogram for test-takers' monologic speaking scores in discourse management

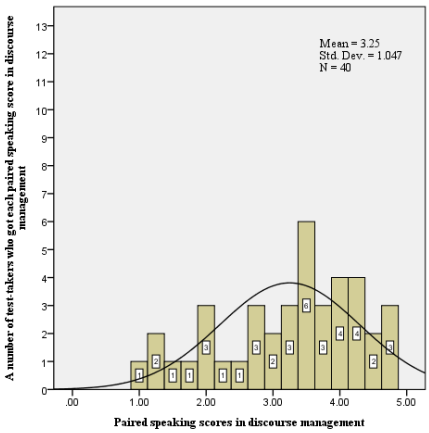


Figure 4.5: Histogram for test-takers' paired speaking scores in discourse management

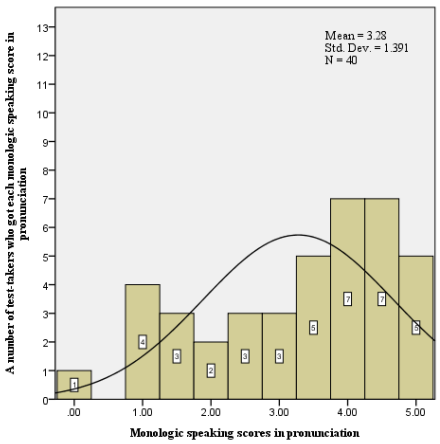


Figure 4.6: Histogram for test-takers' monologic speaking scores in pronunciation

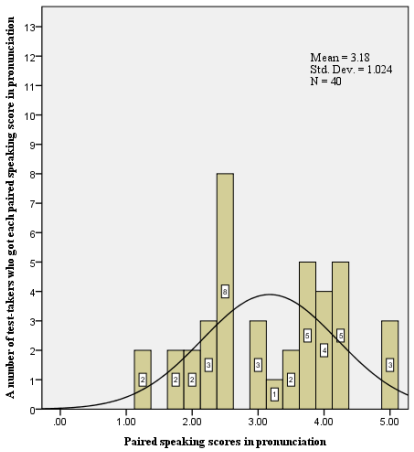


Figure 4.7: Histogram for test-takers' paired speaking scores in pronunciation

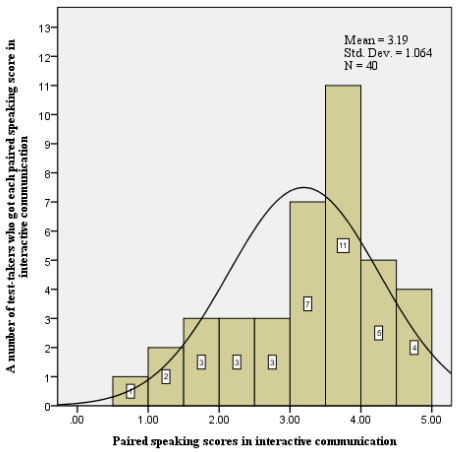


Figure 4.8: Histogram for test-takers' paired speaking scores in interactive communication

The descriptive statistics for the two types of pairing are presented in Table 4.2, which is followed by the histograms for the analytical scores of the paired speaking test in the shared L1 pairs and the non-shared L1 pairs (Figures 4.9–4.16).

Table 4.2: Descriptive statistics for analytical scores in the paired speaking tests with shared and non-shared L1 pairs

Test	Category		Min	Max	Mean	Median	SD
Paired speaking	Grammar and vocabulary (1–5 points)	SL	.50	5.00	3.36	3.50	1.12
		NSL	.50	5.00	3.36	3.50	1.21
	Discourse management (1–5 points)	SL	.50	4.50	3.13	3.50	1.16
		NSL	1.00	5.00	3.45	3.50	1.12
	Pronunciation (1–5 points)	SL	.50	5.00	3.19	3.00	1.16
		NSL	1.00	5.00	3.19	3.00	1.10
	Interactive communication (1–5 points)	SL	.50	5.00	3.23	3.50	1.20
		NSL	.00	5.00	3.16	3.25	1.28

Note: SL refers to shared L1 pairs.
NSL refers to non-shared L1 pairs.
The total possible score for each analytical score for the monologic and paired speaking tests is 5.

The minimum scores of the shared L1 pairs in all categories were 0.50. The maximum score in the grammar and vocabulary (Mean=3.36, SD=1.12), pronunciation (Mean=3.19, SD=1.16) and interactive communication (Mean=3.23, SD=1.20) categories was 5.00 and in the discourse management category (Mean=3.13, SD=1.16) was 4.50.

For the non-shared L1 pairs, the minimum scores in grammar and vocabulary, discourse management, pronunciation and interactive communication were 0.50, 1.00, 1.00 and 0.00, respectively. The maximum score in all categories was 5.00. The mean scores in the grammar and vocabulary, discourse management, pronunciation and interactive communication categories were 3.36 (SD=1.21), 3.45 (SD=1.12), 3.19 (SD=1.10) and 3.16 (SD=1.28), respectively.

The histograms for the paired speaking scores in the shared and non-shared L1 pairs) are presented in Figures 4.9–4.16.

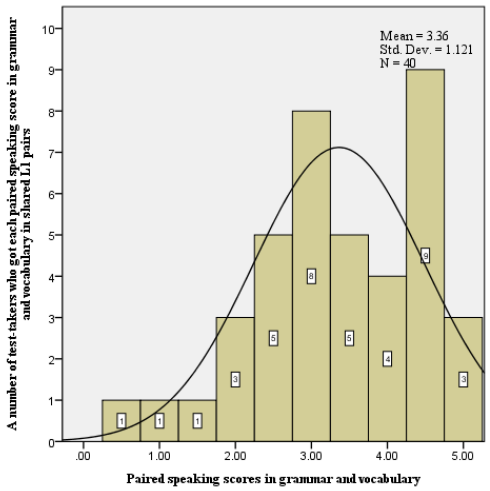


Figure 4.9: Histogram for test-takers' paired speaking scores in grammar and vocabulary in shared L1 pairs

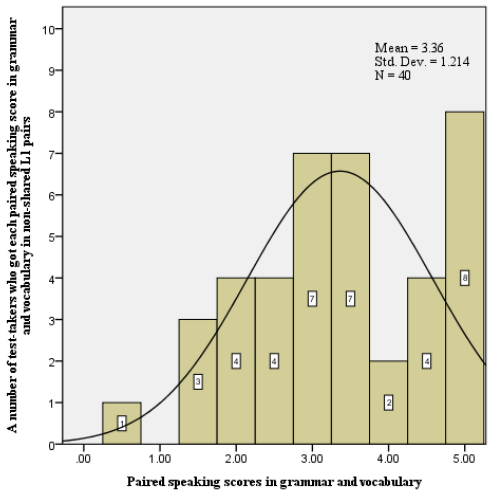


Figure 4.10: Histogram for test-takers' paired speaking scores in grammar and vocabulary in non-shared L1 pairs

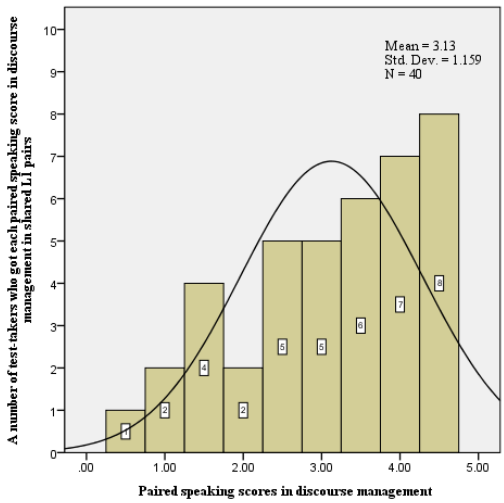


Figure 4.11: Histogram for test-takers' paired speaking scores in discourse management in shared L1 pairs

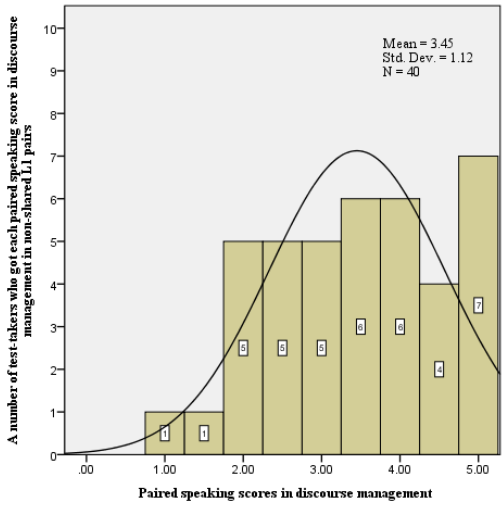


Figure 4.12: Histogram for test-takers' paired speaking scores in discourse management in non-shared L1 pairs

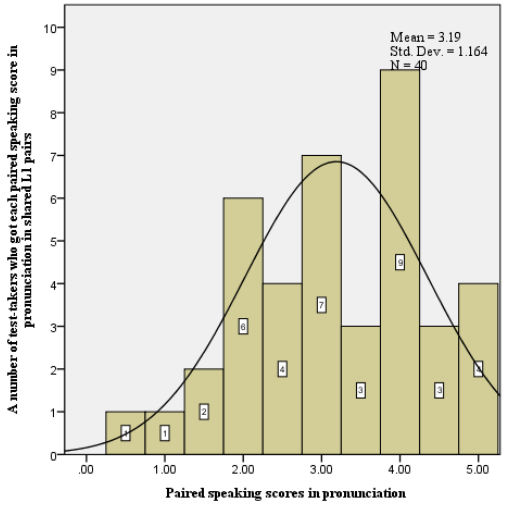


Figure 4.13: Histogram for test-takers' paired speaking scores in pronunciation in shared L1 pairs

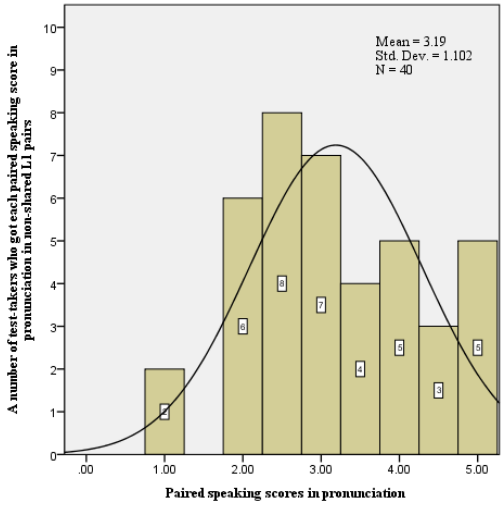


Figure 4.14: Histogram for test-takers' paired speaking scores in pronunciation in non-shared L1 pairs

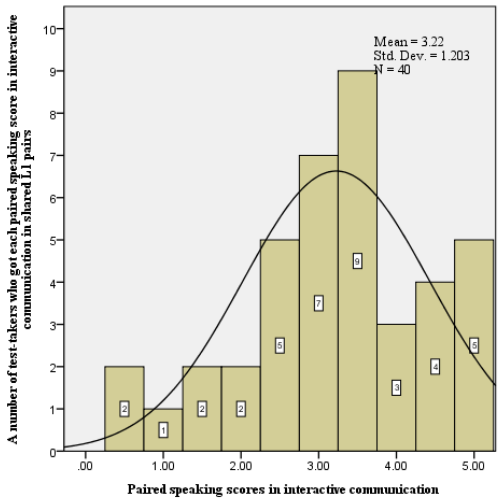


Figure 4.15: Histogram for test-takers' paired speaking scores in interactive communication in shared L1 pairs

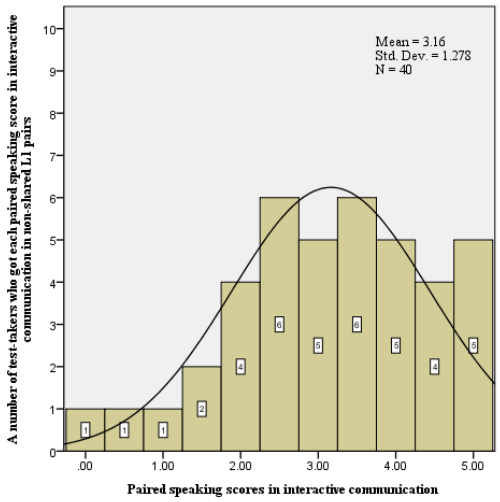


Figure 4.16: Histogram for test-takers' paired speaking scores in interactive communication in non-shared L1 pairs

The histograms for the test-takers' listening and speaking scores (monologic speaking, paired speaking, paired speaking in shared and non-shared L1 pairs) illustrated in Figures 4.1–4.16 all show that the score distributions of those tests were not normal. Therefore, non-parametric tests were used to examine the statistical differences (Wilcoxon Signed Rank Test) and correlations (Spearman's correlation) of the variables in the main study.

4.2 Reliability of the speaking tests

The listening test was pretested to check its reliability, and this was reported as .91 (see Chapter 3, Section 3.4.2.1 for more details). This section will now illustrate the reliability of the two raters in the speaking tests.

As briefly discussed in Section 3.4.4, it was decided to investigate two types of inter-rater reliability, relative reliability and absolute agreement reliability, to assess the degree of inter-rater reliability obtained in the speaking tests, since obtaining both would reflect raters' consistency more accurately (Graham et al., 2012). It should be noted that unlike in the pilot study, a rater training session was provided in the main study in order to enhance consistency in awarding speaking scores by the two raters.

For the relative reliability, the monologic and paired speaking scores in each analytical category (grammar and vocabulary, discourse management, pronunciation and interactive communication) gained from the two raters were tested by using Pearson's correlation. Pearson's correlation measures the correlation among the linear variables (Hinton et al., 2004, p. 297) and therefore measures the relative relationship of the two raters' scores. Nevertheless, the Pearson coefficient considers only a relative order, and even a correlation coefficient of 1.0 does not mean that the raters awarded the same scores. In order to determine the inter-rater absolute agreement, the scores given by the two raters were also examined by calculating the percentage for raters' scoring agreement (Salkind, 2011, p. 108).

Table 4.3 illustrates the results for the relative reliability. It should be noted that there are no data for interactive communication in the monologic speaking test, since the category was not applied to the monologic task. Hence, the total possible score for the monologic speaking task is 15 and the total possible score for paired speaking tasks is 20.

Table 4.3: Inter-rater relative reliability for the monologic and paired speaking tests

	Grammar and vocabulary		Discourse management		Pronunciation		Interactive communication	
	Mono	Pair	Mono	Pair	Mono	Pair	Mono	Pair
Pearson correlation	.93**	.92**	.88**	.91**	.89**	.92**	-	.90**
Sig. (2-tailed)	.00	.00	.00	.00	.00	.00	-	.00

Note: ** Correlation is significant at the 0.01 level (2-tailed).
Mono refers to the monologue speaking test.
Pair refers to the paired speaking tests.

Table 4.3 clearly shows that there are significant positive correlations between the two raters when they are scoring both the monologic and paired speaking tests in each analytical category at the 0.01 level. The Pearson correlation coefficients for the monologic speaking test for grammar and vocabulary, discourse management and pronunciation were 0.93, 0.88 and 0.89, respectively. The Pearson correlation coefficients for the paired speaking tests for grammar and vocabulary, discourse management, pronunciation and interactive communication were .92, .91, .92 and .90, respectively. According to Hinton et al. (2004, p. 364), a reliability coefficient of 0.70 or above shows sufficiently high reliability in performance tests; therefore, it can be considered that the inter-rater reliability when awarding speaking scores was satisfactory.

In order to confirm the actual agreement of the two raters' scores in each analytical category, the inter-rater agreement was investigated by calculating the percentage of exact and adjacent agreement. The results are shown in Table 4.4.

Table 4.4: Percentage of inter-raters' exact and exact and adjacent agreement when scoring monologic and paired speaking tests

Category	Task type	Exact agreement		One point of difference		Exact and adjacent agreement	
		Freq	%	Freq	%	Freq	%
Grammar and vocabulary (1–5 points)	**Mono**	26	65.00	14	35.00	40	100.00
	Pair	40	50.00	40	50.00	80	100.00
Discourse management (1–5 points)	**Mono**	20	50.00	20	50.00	40	100.00
	Pair	44	55.00	36	45.00	80	100.00
Pronunciation (1–5 points)	**Mono**	22	55.00	18	45.00	40	100.00
	Pair	52	65.00	28	35.00	80	100.00
Interactive communication (1–5 points)	**Mono**	-	-	-	-	-	-
	Pair	39	48.75	41	51.25	80	100.00

Note: Freq refers to frequency.
Mono refers to the monologic speaking test.
Pair refers to the paired speaking tests.

The percentages for inter-raters' exact agreement for the grammar and vocabulary, discourse management and pronunciation categories of the monologic speaking task were 65.00, 50.00 and 55.00, respectively. The percentage for inter-raters' exact agreement for the grammar and vocabulary, discourse management, pronunciation and interactive communication categories of the paired speaking tasks were 50.00, 55.00, 65.00 and 48.75, respectively. While the percentage for the exact agreement was not very high, the percentage for the combination of exact and adjacent agreement for all rating categories in both types of test reached 100%. According to Graham et al. (2012), an acceptable percentage for exact and adjacent agreement is around 90%. This therefore suggests that the inter-rater agreement of the two raters in the main study was satisfactory.

Based on the above two types of inter-rater reliability examinations, which indicated that Raters 1 and 2 were acceptably reliable, it was decided to calculate the average scores of the two raters and to treat them as test-takers' speaking scores for each task.

Before RQ1 and RQ2 are addressed, it was thought that it would be interesting to examine how test-takers' paired test performance would compare with their monologic test performance. A non-parametric Wilcoxon Signed Rank Test was

conducted, as shown in Table 4.5.

Table 4.5 Statistical differences between analytical scores in monologic and paired speaking tests (N=40)

Category	Task	Mean	Median	SD	Min	Max	Wilcoxon
Grammar and vocabulary (1–5 points)	Mono	3.58	4.00	1.48	.50	5.00	Z = -1.55 p = .12
	Pair	3.36	3.38	1.09	.75	5.00	
Discourse management (1–5 points)	Mono	3.33	4.00	1.43	.50	5.00	Z= -.68 p = .50
	Pair	3.25	3.50	1.05	1.00	4.75	
Pronunciation (1–5 points)	Mono	3.28	3.50	1.39	.00	5.00	Z= -1.08 p = .28
	Pair	3.18	3.13	1.02	1.25	5.00	
Interactive communication (1–5 points)	Mono	-	-	-	-	-	-
	Pair	3.19	3.50	1.06	.75	5.00	

Note: Mono refers to the monologic speaking test.
Pair refers to the paired speaking tests.

There were no statistically significant differences between the test-takers' scores in the monologic and paired speaking tests in any analytical category at the 0.05 level. This means that, unlike Brooks' (2009) finding that higher scores were awarded to paired tests than to monologic tests, the test-takers of this study did not demonstrate different speaking ability in the two types of test. This might relate to the monologic task used in this study, which was made to be comparable with the paired tasks as much as possible. Although this result is not directly relevant to the research questions of this study, it offers some additional understanding of the test-takers' performance in both speaking formats, whose scores will be further analysed in the next section.

4.3 Quantitative results for Research Questions 1 and 2

Having described the test-takers' listening and speaking scores, this chapter now moves on to present the results of the quantitative analyses that are relevant to Research Questions 1 and 2.

> **RQ1:** *To what extent is test-takers' performance in paired speaking tests in shared and non-shared L1 pairs affected by their listening proficiency?*
>
> **RQ2:** *Are there any differences in speaking scores when test-takers are paired with shared L1 partners as compared to (when they are paired with) non-*

shared L1 partners?

4.3.1 Relationship between listening scores and speaking scores in monologic and paired speaking tests in both types of pair (RQ1)

As confirmed in Section 4.1 (Figures 4.1–4.16), since both listening and speaking scores were not normally distributed, Spearman correlation was employed to discover the relationship between the following:

- listening scores and monologic speaking scores in each analytical category; and
- listening scores and paired speaking scores in each analytical category with (a) both shared and non-shared L1 pairs together, (b) shared L1 pairs, and (c) non-shared L1 pairs.

A summary of the correlations for the monologic and paired tests for all participants is presented in Table 4.6.

Table 4.6: Correlation between listening scores and analytical scores of monologic and paired speaking tests (N=40)

	Grammar and vocabulary		Discourse management		Pronunciation		Interactive communication	
	Mono	**Pair**	**Mono**	**Pair**	**Mono**	**Pair**	**Mono**	**Pair**
Spearman's rho	.19	.32*	.13	.35*	.19	.25	-	.08
Sig. (2-tailed)	.25	.04	.44	.03	.24	.13	-	.63

Note: Mono refers to the monologic speaking test.
Pair refers to the paired speaking tests.

Table 4.6 shows that there was no statistically significant correlation between the listening scores and the monologic speaking scores in each analytical category.

For the correlations between the listening scores and the paired speaking scores, there were statistically significant correlations between the following:

- the listening scores and the grammar and vocabulary scores ($r_s = .32$, $p = .04$); and
- the listening scores and the discourse management scores ($r_s = .35$, $p = .03$) at the .05 level.

The strength of the correlations was moderate for both cases, according to Cohen's

(2009) interpretation of correlation coefficients. The former result indicates that the better test-takers' listening was, the more accurately and appropriately they used grammar and vocabulary. The finding may not be very surprising, as grammar and vocabulary elements usually account for a significant amount of the total score variance in skill-specific tests (e.g., Geranpayeh, 2007; Shiotsu and Weir, 2007). A significant correlation was also found in the pilot study between the listening scores and grammar and vocabulary scores in the monologic speaking test (see Section 3.3.5.1).

Of interest to this study is the statistically significant, positive correlation between the listening scores and the discourse management scores in the paired speaking tests. This result is congruent with that of the pilot study. It shows that the better the test-takers' listening was, the more effectively they managed discourse in the paired speaking tests. Intuitively, it makes sense that when the test-takers understand their partner's speech better, they can speak more coherently, relating their output to their partner's utterances. They may not have to hesitate due to comprehension problems, which could result in more fluent discourse.

Table 4.7 now compares the correlations of the listening and paired speaking scores between shared and non-shared L1 pairs.

Table 4.7: Correlation between listening scores and analytical speaking scores for shared and non-shared L1 pairs (N=40)

	Grammar and vocabulary		**Discourse management**		**Pronunciation**		**Interactive communication**	
	shared L1	**non-shared L1**	**shared L1**	**non-shared L1**	**shared L1**	**non-shared L1**	**shared L1**	**non-shared L1**
Spearman's rho	.26	.37*	.26	.38*	.22	.22	-.01	.14
Sig. (2-tailed)	.10	.02	.11	.02	.17	.18	.97	.37

Note: *Correlation is significant at the 0.05 level (2-tailed).

Interestingly, only the non-shared L1 pairs showed a statistically significant correlation between the following:

- test-takers' listening scores and grammar and vocabulary scores (r_s*=.37, p=.02*); and

- test-takers' listening scores and discourse management scores ($r_s = .38$, $p = .02$).

These are exactly the same two categories in which correlations with listening scores were reported for the entire group. Therefore, the correlations shown in Table 4.6 actually reflected the correlations found in the non-shared L1 pairs presented in Table 4.7. While the strength of the two correlations was again only moderate, the positive relationship between the test-takers' listening scores and discourse management scores in pairs applies only to the non-shared L1 pairs, not the shared L1 pairs. This means that when the test-takers who were paired with a non-shared L1 partner had good listening proficiency, they could comprehend their partner and manage the discourse better, which means, according to the Cambridge First rating scale (UCLES, 2015), producing more extended stretches of language, providing more relevant and clearer organisation of ideas, and using more cohesive devices and discourse markers. In contrast, the shared L1 pairs might not have required as much English listening proficiency to understand their partner. This suggests that even when the test-takers lacked listening skills, they might have understood their shared L1 partners to the extent that they could manage the discourse as well as those who had better listening skills.

The results of the main study did not confirm the pilot results presented in Section 3.3.5.1, which indicated a statistically significant correlation between the listening and (a) discourse management scores only in shared L1 pairs, and (b) pronunciation and interactive communication scores only in non-shared L1 pairs. These contradictory results between the pilot and main studies might relate to the differences in the numbers and L1s of the participants, but more importantly to the improved listening test used in the main study, and more controlled test-taker variables (e.g., age and gender) in the main study. It was therefore hoped that the main study results would be a more accurate reflection of the impact of test-takers' listening ability on paired test scores in shared and non-shared L1 pairs.

Table 4.8: Statistic differences between the analytical scores in paired speaking tests in shared and non-shared L1 pairs (N=40)

Category	Type	Mean	Med	SD	Min	Max	Wilcoxon
Grammar and vocabulary (1–5 points)	shared L1	3.36	3.50	1.12	.50	5.00	Z = -.12 p = .90
	non-shared L1	3.36	3.50	1.21	.50	5.00	
Discourse management (1–5 points)	shared L1	3.13	3.50	1.16	.50	4.50	Z = -1.90 p = .06
	non-shared L1	3.45	3.50	1.12	1.00	5.00	
Pronunciation (1–5 points)	shared L1	3.19	3.00	1.16	.50	5.00	Z = -.23 p = .82
	non-shared L1	3.19	3.00	1.10	1.00	5.00	
Interactive communication (1–5 points)	shared L1	3.23	3.50	1.20	.50	5.00	Z = -.81 p = .86
	non-shared L1	3.16	3.25	1.28	.00	5.00	

Table 4.8 indicates that there was no statistically significant difference for each analytical score in the two types of pairing at the 0.05 level. This is in line with the pilot study results reported in Section 3.3.5.2. It seems that types of pairing do not affect the test-takers' paired test scores.

4.4 Summary of the main findings of the quantitative analyses

This chapter has firstly reported the descriptive statistics for the listening scores and the speaking scores and the reliability of the speaking ratings. It then examined the relationship between listening and the speaking scores in both the monologic and paired speaking tests with the entire group, with shared and non-shared L1 pairs (RQ1). It also compared paired speaking scores awarded to the shared and non-shared L1 pairs (RQ2). The main findings in relation to RQs 1 and 2 will be summarised below.

As for RQ1, the main findings include the following:

- there was no statistically significant correlation between the test-takers' listening and monologic speaking scores in any category;
- there was a statistically significant positive correlation between the test-takers' listening scores and paired speaking scores for grammar and vocabulary (r_s=*.32, p=.04*) for the entire group, but this in fact reflected the significant correlation only for the non-shared L1 pairs (r_s=*.37, p=.02*); and

- there was a statistically significant positive correlation between the test-takers' listening scores and paired speaking scores for discourse management (r_s=*.35, p=.03*) for the entire group, but this in fact reflected the significant correlation only for the non-shared L1 pairs (r_s=*.38, p=.02*).

All correlation coefficients reported were moderate, according to Cohen's (1988) definition (i.e., small: r=0.1, moderate: r=0.3, large: r=0.5).

The above results suggest that test-takers' listening proficiency did not matter in shared L1 pairs. While the result for grammar and vocabulary scores is not very surprising (see Section 4.3 for the explanation), what is interesting is that the greater the listening proficiency test-takers had, the more effectively they managed their discourse. Although the strength of the correlations was only moderate, listening proficiency was more important for the test-takers' speaking performance in the non-shared L1 pairs than in the shared L1 pairs. It should, however, be acknowledged that the skill profile for test-takers with shared L1s might be more similar, regardless of their overall performance, which might have confounded the results.

Regarding RQ2, the results indicated that

- there was no statistical difference in the paired speaking scores that the test-takers received when they were paired with a shared L1 partner as compared to when they were paired with a non-shared L1 partner.

This suggests that the test-takers' paired speaking scores were not significantly affected by the types of pairing (i.e., with shared or non-shared L1 pairs).

This chapter has reported the qualitative results for RQ1 and RQ2; the next chapter will present and discuss the results from the interactional data analysis, while attempting to interpret and elaborate on the quantitative results presented in this chapter.

CHAPTER 5 Results of Interactional Data Analysis

This chapter presents the results of the interactional data analysis together with the analysis of test-takers' and raters' stimulated recall interviews. In doing so, it also attempts to interpret and elaborate on the score findings presented in Chapter 4.

Firstly, similarities and differences in communication patterns related to interactive listening between the shared and non-shared L1 pairs are described, while illustrating and discussing relevant excerpts from their spoken data (Section 5.1). This is to address RQ3 *(What are the similarities and differences in communication patterns between shared and non-shared L1 pairs?)*. Wherever appropriate, test-takers' stimulated recall interview data are provided to support the interpretation of the interactional data. Secondly, some additional features observed from interactions which are not necessarily related to interactional listening but are still relevant in discussing the differences between the shared and non-shared L1 pairs are reported (Section 5.2). Thirdly, results from examiners' stimulated recall interviews are described in order to illuminate the results from both score and interactional data analyses (Section 5.3). Finally, Section 5.4 presents and discusses the summary of the main findings.

5.1 Communication patterns related to interactive listening between shared and non-shared L1 pairs (RQ3)

The following sections present the findings of CA on test-takers' paired speaking performance. Explanations, supporting excerpts and quantification of each category are portrayed. The analysis procedures and coding scheme are described in Section 3.4.4.

5.1.1 Similarities in communication patterns related to interactive listening between the shared and non-shared L1 pairs

Communication patterns related to interactive listening that were found to be similar between the two types of pair relate to the following:

- supplying relevant vocabulary;
- demonstrating comprehension; and
- back-channelling.

An explanation of each communication pattern, including supporting excerpts and quantification, is presented in the following sub-sections. The supporting excerpts of non-shared L1 pairs are illustrated first and are followed by excerpts of shared L1 pairs.

5.1.1.1 Supplying relevant vocabulary

Supplying appropriate vocabulary occurs when a test-taker provides a word/phrase which his/her partner is searching for. It shows that the test-takers are paying enough attention and have sufficient comprehension to predict a missing word, which helps the interaction to continue (Ducasse, 2010). The CA results of the current study indicated that supplying vocabulary was utilised by the test-takers in both types of pair. The listener test-takers provided a word that their partner was searching for to complete their partner's utterance. It was usually triggered when the test-taker as a listener felt that their partner was searching for a word/phrase. By supplying a relevant word/phrase, a listener was able to show his/her engagement and demonstrate comprehension.

The quantitative comparison of the number for supplying relevant vocabulary in the shared L1 pairs and the non-shared L1 pairs is presented first, followed by examples to show the qualitative similarities of such occurrences. The frequencies for supplying relevant vocabulary of the test-takers in the shared and non-shared L1 pairs are shown in Table 5.1.

Table 5.1: Frequencies for supplying relevant vocabulary of test-takers in shared and non-shared L1 pairs

Types of pair	Pairing	Frequency	Percentage
Shared L1	Urdu–Urdu	8	11.77
	Thai–Thai	25	36.76
	Total	33	48.53
Non-shared L1	Urdu–Thai	35	51.47
Total		68	100.00

As shown in Table 5.1, the frequencies for supplying relevant vocabulary of

the test-takers in both types of pair were similar. The shared L1 pairs supplied relevant vocabulary 33 times (48.53%), while the non-shared pairs supplied relevant vocabulary 35 times (51.47%). However, the numbers for supplying relevant vocabulary in Urdu and Urdu L1 pairs and those for the Thai and Thai L1 pairs were very different. While Thai L1 pairs supplied relevant vocabulary 25 times, Urdu L1 pairs did it only 8 times. This huge difference between Thai L1 pairs and Urdu L1 pairs indicates that this communication pattern might be associated more with Thai L1 speakers than Urdu L1 speakers. The inter-coder agreement for the shared L1 pairs was .85 and for the non-shared L1 pairs it was .83, both indicating that the level of coding reliability was acceptable.

A non-parametric Wilcoxon Signed Rank Test was utilised to investigate the statistical difference in supplying vocabulary of the test-takers in the two types of pair.

Table 5.2: Statistical difference for filling a silence of test-takers in shared and non-shared L1 pairs

Interactive listening clue	Type	Mean	Med	SD	Min	Max	Wilcoxon
Filling a silence	**SL**	.83	.00	1.15	.00	4.00	Z = -.56
	NSL	.88	1.00	1.22	.00	5.00	*p* = .58

Note: SL refers to shared L1 pairs.
NSL refers to non-shared L1 pairs.

As shown in Table 5.2, there was no statistically significant difference between the test-takers in the shared and non-shared L1 pairs regarding filling the silence by supplying vocabulary. This means that the test-takers in both types of pair were similarly supplying vocabulary when their partner was searching for a word or trying to deliver an idea.

Excerpt 6 illustrates a conversation between a non-shared L1 pair. P11 was trying to describe a variety of professions on the prompt card but her talk was very hesitant, as shown by a number of filled and unfilled pauses. In Line 15, when she attempted to select one profession to focus on further, she could not think of a word to explain it. Her partner (T11) recognised the difficulty from her inhaling and a filled pause of "*er*", and assisted P11 by supplying the word "*popular*" (Line 16). P11 accepted T11's assistance by uttering a back-channelling "*yeah*" and continued

presenting her idea.

Excerpt 6

Topic: Professions (P11=Pakistani female 11, T11=Thai female 11)

L06	P11:	ah::.hhh (.5) here ((pointing at a picture)) is a lot of er:
L07		mm::it is just like ah:: that is like a model ah:: mm:: er::
L08		(0.3) this one is like a painting
L09	T11:	uh huh
L10	P11:	scientist singing
L11	T11:	yeah
L12	P11:	((clearing throat)) yeah so everyone has a lot of
L13		advantages and disadvantages in these careers
L14	T11:	uh huh
L15	P11:	but in the uk i think this is this one is .hh er:::
→ L16	T11:	Popular
L17	P11:	yeah .hhh this one is (.) more important

In the stimulated recall interview with T11, she revealed, "*My partner couldn't think of a word to explain her idea. So I said 'popular' cos I thought it was the word that she was looking for.*"

Supplying a vocabulary was also observed in Thai L1 pairs. Excerpt 7 is part of a conversation between Thai L1 test-takers. They were discussing the topic of professions. T17 was trying to explain what qualifications might be necessary to become an artist, but she struggled to come up with a relevant word, as indicated by a hand gesture and a short pause followed by a filler, "*I mean*". Latching onto the filler, T18 assisted her by offering the word "*talent*", as shown in Line 54. T17 accepted the word "*talent*" by saying "*yeah the talent*" (Line 55). The assistance of T18 helped T17 to continue explaining her idea.

Excerpt 7

Topic: Professions (T17=Thai female 17, T18=Thai female 18)

L52	T17:	ah::: it seems to be er::: easy work but actually it is not
L53		if you don't have the .hh ((moving hand)) (.) i mean=
→ L54	T18:	=talent yeah↑
L55	T17:	yeah↓ the talent cos you need to make to picture
L56		((moving hands))

When T18 was interviewed after finishing conversing with T17 about why she uttered the word "*talent*", she said, "*I was listening to my partner. I think…er…I thought it's the word she wanted to say, so I spoke it out*."

Excerpt 8 shows a conversation between two Urdu L1 test-takers. P20 was talking about good and bad friends. When she talked about bad friends, she paused for a while and repeated "*they will*" twice while searching for a word to describe them. P19 assisted her by saying "*destroy*", as shown in Line 46, and she agreed to use the given word. This helped the conversation continue.

Excerpt 8

Topic: Friends (P19 = Pakistani female 19, P20 = Pakistani female 20)

	L40	P20:	can advise (.) you in good manner and so they can
	L41		guide you on (.) right thing ((raising hands)) .hh (0.3)
	L42		so (.) you can learn .hh and er: (0.5) er: follow them
	L43		(0.3) er: to obtain your objective (.) in right feet .hh
	L44		otherwise if you have bad friends (.) they will they
	L45		will [er: they=
→	L46	P19:	[destroy ha ha ha
	L47	P20:	=will destroy your life they will they will disappoint
	L48		you

In stimulated recall interview with P19, she reported, "*[M]y partner was talking about having bad friends. I thought that having bad friends might give negative effects on our lives, so I said 'destroy'. My partner seemed to agree with me because she used the word I had said."* P20 reported, "*I was trying to find the best word to describe the effect of having bad friends and then my partner helped me by saying that word. You know... er… it was an exact word I was looking for.*"

As such, showing interactive listening through supplying relevant vocabulary was found in both types of pair. It not only demonstrated the test-takers' comprehension of their partner's talk but also assisted the conversation to continue smoothly.

5.1.1.2 Demonstrating comprehension

Demonstrating comprehension can be evidenced from the test-takers' comments about a partner's contribution (Ducasse, 2010). It can be examined by checking whether it coherently relates to the partner's message and whether the test-taker can respond relevantly to their partner's question. However, short answers (e.g., "*yes*", "*no*" or "*I agree with you*") were not considered as demonstrating comprehension in the current study, because the test-takers might have been able to provide such

short answers without fully comprehending what the partner said, and it is not possible to code these short responses reliably.

The frequencies for demonstrating comprehension of the test-takers in the shared L1 pairs and the non-shared L1 pairs are shown in Table 5.3.

Table 5.3: Frequencies for demonstrating comprehension of test-takers in shared and non-shared L1 pairs

Types of pair	Pairing	Frequency	Percentage
Shared L1	Urdu–Urdu	56	19.58
	Thai–Thai	78	27.27
	Total	134	46.85
Non-shared L1	Urdu–Thai	152	53.15
Total		286	100.00

As shown in Table 5.3, the test-takers in both types of pairs demonstrated comprehension a total of 286 times. The frequencies for demonstrating comprehension of the test-takers in the shared and the non-shared L1 pairs were 134 (46.85%) and 152 (53.15%), respectively. Urdu and Urdu L1 pairs demonstrated comprehension 56 times (19.58%), while Thai and Thai L1 test-taker pairs demonstrated comprehension 78 times (27.27%). The inter-coder agreement calculated for this coding was .71 for the shared L1 pairs and .70 for the non-shared L1 pairs, indicating that the coding reliability for both types of pair were acceptable.

The quantitative data for demonstrating comprehension in both types of pair was analysed by a non-parametric Wilcoxon Signed Rank Test. As illustrated in Table 5.4, there was no statistically significant difference in the communicative occurrences that demonstrated comprehension between the test-takers in the shared and non-shared L1 pairs.

Table 5.4: Statistics for demonstrating comprehension of test-takers in shared and non-shared L1 pairs

Interactive listening clue	Type	Mean	Med	SD	Min	Max	Wilcoxon
Demonstrating comprehension	SL	3.35	3.50	2.01	.00	8.00	Z = -.66
	NSL	3.80	3.00	2.74	.00	11.00	p = .51

Note: SL refers to shared L1 pairs.
NSL refers to non-shared L1 pairs.

An investigation of the interactional data suggested that the test-takers in both

the shared and non-shared L1 pairs demonstrated comprehension to present the evidence of their interactive listening during a conversation. Excerpt 9 shows a conversation between Thai L1 and Urdu L1 test-takers. They talked about the topic of professions. T03 explained that one profession would require some specific qualifications, one of which was then relevantly followed up by P03, demonstrating P03's comprehension of what T03 had said, and he added his own idea, as shown in Line 57.

Excerpt 9

Topic: Professions (T03=Thai male 03, P03=Pakistani male 03)

L51	T03:	ah:: (0.3) left side the most difficult .hhh they- they-
L52		they need a skill, (.) talent and er some (0.5) more
L53		motivation ((raising hand)) for this uh huh::
L54		businessman, they need money but it's more- more
L55		more creative, more motivation ((looking at partner
L56		and nodding head))
→ L57	P03:	yeah ((nodding head)) motivation is important because
L58		without motivation, you can't achieve anything ah:
L59		mm:: ((clearing throat))

In stimulated recall interview, P03 reported, "*I said 'yeah' and nodded my head because I agreed with him about the qualifications of a businessman especially motivation. I think it's important for achievement in every profession.*"

Excerpt 10 shows the demonstrating comprehension between the test-takers in the shared L1 pair. Thai L1 test-takers were discussing professions. T09 gave examples of a successful football player when suggesting the importance of talent in order to be successful in football. T10 ratified the topic, supporting T09's idea by referring to the name of the football player mentioned by T09, as shown in Line 27. This clearly demonstrated T10's understanding of T09's speech.

Excerpt 10

Topic: Professions (T09=Thai male 09, T10=Thai male 10)

L24	T09:	they got like so much talent. they got like Ronaldo or
L25		Ronaldinho what- what do you think? .hh they have to
L26		practice a lot or:::
→ L27	T10:	yes actually (.) it's ah:: (0.5) if you talk about Ronaldo
L28		or Messi or some people like this. i think (0.3) .hh you
L29		can say (0.5) they are (0.3) born to be a football player
L30	T09:	ok uh huh

In the stimulated recall interview with T10, he reported, "*I agreed with my partner. The famous football players like Ronaldo and Messi are talented. And I think they are born to be the football player*."

Next is an example of demonstrating comprehension between a pair of Urdu L1 test-takers. In Excerpt 11, it was evidenced through responding relevantly to their partner's question and extending the partner's idea. When P12 asked P11 the question "*What do you think?*" (Line 05), P11 responded relevantly by comparing relationships between friends and between family members (Line 06). P12 seemed to agree with P11's idea and she extended and supported P11's idea that people could share everything with friends, while they might not do so with their family members (Line 16).

Excerpt 11

Topic: Friends (P11=Pakistani female 11, P12=Pakistani female 12)

	L05	P12:	what do you think?
→	L06	P11:	(0.3) ah::: the advantage of having friends (.) er:: is ah::
	L07		you know (.) the friendship is a .hhh very (0.5) good
	L08		((raising hands)) relation (0.4) rather than (.) ah::
	L09		instead of (.) brothers, sisters .hh ah:: (.) you know er: if
	L10		you have a friend you discuss your feeling [you::=
	L11	P12:	[yeah
	L12	P11:	=discuss <u>ev</u>erything (.) like you don't discuss with your
	L13		.hh mm::: (.) very close relatives, sister [or brothers=
	L14	P12:	[yeah yeah
	L15	P11:	=.hh and er:: like like [a husband ha ha ha
→	L16	P12:	[yes oh yes they-they are really
	L17		close because you can share (0.3) ah::: (.) anything .hhh
	L29		but even we can't say with our parents
	L30		[we can share them everything ((nodding head))
	L30	P11:	[ah yeah heh heh heh

As shown in the excerpts above, relevantly responding to or appropriately developing what his/her partner has said demonstrates a listener's interactive listening and comprehension. If clarification requests are made, this illustrates the test-takers' incomprehension. When such requests are used to negotiate meaning until the listener fully understands the speaker's message and the listener is ready to change their role from listener to speaker, it shows the test-taker's comprehension. If clarification questions are not asked when necessary, or if the speaker cannot manage to respond to such questions, the lack of clarity or ambiguity

can cause communication breakdowns.

5.1.1.3 Back-channelling

Back-channelling is a universal feature of human communication, and it is a part of conversation which a listener performs (Shelly and Gonzalez, 2013). Back-channelling is defined as a speech sound which is produced by an interactive listener in order to provide supporting feedback to a speaker while the speaker maintains the floor (Ducasse, 2010; Ducasse and Brown, 2009). It is used by a listener to let a speaker know that he/she is listening to and understanding what the speaker is saying. There are two sorts of back-channelling: verbal (e.g., *yeah*, *ok*, *uh huh* and *mm*) and non-verbal, e.g., nodding the head (Ducasse, 2010; Ducasse and Brown, 2009; Shelly and Gonzalez, 2013). However, Ducasse (2010) did not include other non-verbal signals in her study, for example, gaze and gesture, laughter, body position and facial expression as part of interactive listening. Those signals were categorised as interpersonal non-verbal communication, which was one of the three categories (interpersonal non-verbal communication, interactive listening and interactional management) for achieving interactive communication in pairs. The current study followed the categorisation used in Ducasse's study, and did not consider other non-verbal signals, except nodding the head, as back-channelling. This was also due to the difficulty of analysing other non-verbal signals reliably.

In this study, there was evidence that the test-takers in both the shared and non-shared L1 pairs used back-channelling while they conversed. The frequencies for back-channelling use of the test-takers in the shared (Urdu–Urdu, Thai–Thai) and non-shared L1 pairs are shown in Table 5.5.

Table 5.5: Frequencies for back-channelling use of test-takers in shared and non-shared L1 pairs

Types of pair	Pairing	Frequency	Percentage
Shared L1	Urdu–Urdu	118	18.02
	Thai–Thai	140	21.37
	Total	258	39.39
Non-shared L1	Urdu–Thai	397	60.61
Total		655	100.00

The frequency for back-channelling use of the test-takers in the shared and non-shared L1 was 258 (39.39%) and 397 (60.61%), respectively. The test-takers in the non-shared L1 pairs used back-channelling more frequently than the shared L1 pairs. The pairs of Thai L1 test-takers used back-channelling more frequently than Urdu L1 pairs. The Urdu L1 test-taker pairs used back-channelling 118 times (18.02%), while the Thai L1 test-taker pairs used it 140 times (21.37%). The inter-coder reliability rates for this coding in the shared and non-shared L1 pairs were .83 and .84, respectively. Hence, the inter-coder reliability in each pair type was acceptable.

A non-parametric Wilcoxon Signed Rank Test was used to investigate the frequency of back-channelling between the test-takers in the shared and non-shared L1 pairs.

Table 5.6: Statistics for back-channelling use of test-takers in shared and non-shared L1 pairs

Interactive listening clue	Type	Mean	Med	SD	Min	Max	Wilcoxon
Back-channelling	SL	6.45	5.00	4.73	1.00	21.00	Z = -2.86
	NSL	9.93	9.00	6.15	.00	26.00	$p < .01$

Note: SL refers to shared L1 pairs.
NSL refers to non-shared L1 pairs.

Table 5.6 shows that there was a statistically significant difference in back-channelling use between the shared and non-shared L1 pairs. The test-takers in the non-shared L1 pairs used back-channelling more frequently than the shared L1 pairs (Mean = 9.93 and 6.45, respectively). It might be assumed that the non-shared L1 pairs used back-channelling more often than the shared L1 pairs because they realised that they had used their "*listening noises*" to signal their comprehension or support a speaker test-taker to keep speaking until they understood and they could show their comprehension by changing from the listener role to the speaker role (Ducasse, 2010). However, although the difference in the numbers of back-channelling uses between the two types of pair should be noted, the use of back-channelling can still be considered as a similarity between the two pair types, since it was frequently observed in both pair types.

It was found that the test-takers in both types of pair used back-channelling to encourage their partner to continue speaking, as shown in Excerpts 12–14. Excerpt 12 is an example of back-channelling use between the test-takers in the non-shared L1 pairs. Thai L1 and Urdu L1 test-takers were talking about the topic of friends. While P01 was presenting his idea, T01 said "*yes*" with laughter and "*yeah yeah*", as shown in Lines 13 and 18, to illustrate that he was listening to P1 and was supporting him to continue speaking.

Excerpt 12

Topic: Friends (T01=Thai male 01, P01=Pakistani male 01)

L11	P01:	even they're helping in an exam (0.3) while we're
L12		sitting together (0.5) he can help us [even we don't do=
→ L13	T01:	[yes ha ha ha
L14	P01:	=it we always take (0.5) er: he's doing the exam he or
L15		she .hh we can say that [first to do my exam ha ha ha
L16	T01:	[ha ha ha
L17	P01:	[that's the friendship]
→ L18	T01:	[yeah yeah] or we can
L19		go (0.5) to travelling with friends so:: .hh

In T01's stimulated recall interview, he was asked why he uttered "*yes*" and "*yeah yeah*" when listening to his partner speaking. T01 answered, "*I said 'yes' and 'yeah yeah' because I got what my partner said. I always say these words when I'm listening to someone speaking.*" The data from the interview with T01 confirms that he used back-channelling to show his engagement with and understanding of his partner's speech. An example of utilising back-channelling in a conversation of a Thai L1 pair is illustrated in the following paragraph.

Thai L1 test-takers were conversing about friends. T07 answered T08's question about the number of his close friends. While T07 said that he had five or six friends, T08 listened and used back-channelling in the form of "*yeah yeah*" (Line 90) to show his interactive listening and to encourage T08 to continue speaking.

Excerpt 13

Topic: Friends (T07=Thai male 07, T08=Thai male 08)

L88	T08:	ah:: yeah how- how many close friends do you have?
L89	T07:	i think (0.3) about five or six (0.5) friends
→L90	T08:	yeah yeah
L91	T07:	i think i have close friends and that (.) both i can
L92		tell him ev- everything

In the stimulated recall interview with T08, he was asked why he said "*yeah yeah*" while listening to his partner. He reported, "*I wanted to let my friend know that I was listening to him and understood what he had said.*" The next excerpt is a part of a conversation between two Urdu L1 test-takers.

Excerpt 14 exhibits how a listener (P02) demonstrated his interactive listening and supported his partner (P01) to speak by producing back-channelling. P02 uttered back-channelling in the form of "*mm*" (Line 08) and sometimes accompanied it with a gesture, i.e., "*yeah*" with head nodding (Line 10). These actions enabled the speaker to continue explaining his idea.

Excerpt 14

Topic: Professions (P01=Pakistani male 01, P02=Pakistani male 02)

L01	P01:	how difficult is to be successful in these professions like
L02		ah playing football .hhh and::: singing, painting, and
L03		dancing and er: the this doctor ((pointing at a picture))
L04		mm: having a doctoring degree .hh like er:: as a sport
L05		professional we will go for sport first (.) like sampling
L06		football how difficult these are to become .h come in this
L07		and have a success in this sport? hh (0.3)
→L08	P02:	mm::
L09	P01:	like when we don't get a good coach (0.7)
→L10	P02:	yeah ((nodding head))
L11	P01:	we can't get success in this (0.5) football match we can't
L12		(.) get something new (0.3)

An interesting point was found in Excerpt 14. P02 used back-channelling when his partner paused for a while, as shown in Lines 8 and 10. In the stimulated recall interview with P02, he said, "*I wanted my friend to keep speaking. So I filled silence when he stopped speaking. I understood what he said. While I was listening to him, I was also thinking of how to express my idea*." Although P02 seemed not

to pay much attention to listening to P01, he did understand what P01 was saying. P02 used back-channelling only when P01 paused speaking to encourage P01 to continue speaking while P02 was thinking of what to say next when he got a turn. This finding is congruent with what Ducasse (2010) found in her study: a listener not only listens to a speaker, but he or she is also thinking of what to say next.

5.1.2 Differences in communication patterns related to interactive listening between shared and non-shared L1 pairs

While both the shared and non-shared L1 pairs showed several similarities in communication patterns related to interactive listening, there were also some differences, in terms of (1) causing communication problems, (2) understanding unclear utterances and incorrect word use and (3) misunderstanding because of having different cultural backgrounds. Examples with an explanation of the differences are presented in the following sections.

5.1.2.1 Causing communication problems

It is not surprising that some communication problems occur during interactive communication between two non-native speakers (NNS–NNS) test-takers (e.g., Gass and Varonis, 1991). The communication problems found in the main study seemed to relate to both speaker and listener's limited linguistic ability, as well as the effect of their L1 backgrounds. However, only the communication problems related to interactive listening are reported here. Two types of communication problems, which were called 'non-engagement' and 'miscommunication' by Gass and Varonis (1991), were identified in the main study. To provide an overview of the communication problems and the number of solved communication problems in two types of pair, those problems are presented first, followed by examples. Then the details of the difference in communication pattern in terms of attempts to understand a partner completely are discussed.

Hahn and Watts (2011) state that when a communication problem occurs, learners usually attempt to solve it in order to achieve their interactive communication goal. They use various explicit strategies, e.g., clarification requests and body language, to repair the miscommunication in order to achieve their

communication goal. They may also exploit their background knowledge, together with personal and social awareness, to solve the miscommunication.

However, the analysis of communication problems in the shared and non-shared L1 pairs revealed some differences between the two groups. While the test-takers in the shared L1 pairs always solved all communication problems by using various strategies, the non-shared L1 pairs did not always succeed in solving the problems. The types and numbers of communication problems and how the problems were solved in the shared L1 pairs and the non-shared L1 pairs are illustrated in Table 5.7.

Table 5.7: Types, numbers and solutions of communication problems in shared and non-shared L1 pairs

Types of communication problem	Number of communication problems		Number of solved communication problems		How communication problems were solved	
	SL	NSL	SL	NSL	SL	NSL
1. Non-engagement, e.g., being quiet, back-channelling (responding to a Wh-question), and shift to a new topic	4 (28.57%)	10 (71.43%)	4 (100%)	7 (70%)	1. Clarified a question by providing an example 2. Encouraged their partner to answer, e.g., *What do you think*?	1. Repeated a question 2. Encouraged their partner to answer, e.g., *Uh huh*, *You know about it*?, *What do you think*? 3. Paraphrased a question 4. Repeated a keyword in a question
2.Miscommunication, e.g., providing irrelevant response, partly relevant response, clarification request, and confirmation check	14 (32.56%)	29 (67.44%)	14 (100%)	20 (40.82%)	1. Clarification requested, e.g., *What do you mean*?, *What?* 2. Confirmation check, e.g., *You mean …*, S*o you think …*, *So you say …* 3. Responded explicitly when realised their partner's misunderstanding, e.g., *No no …*	1. Clarification requested e.g., *What*?, *Pardon*?, *What is …*?, *Sorry*?, *Where*?, *What do you mean by …*? 2. Confirmation check, e.g., *You mean …*? 3. Paraphrased a word or question 4. Repeated a keyword
Total	18	39	18	27	-	-

Note: SL refers to shared L1 pairs.
NSL refers to non-shared L1 pairs.

As shown in Table 5.7, there were 18 communication problems (4 non-engagements and 14 miscommunications) which occurred in the shared L1 pairs and 39 communication problems (10 non-engagements and 29 miscommunications) in the non-shared L1 pairs. Interestingly, all communication problems in the shared L1 pairs were solved (100%), while only 27 out of the 39 problems (69.23%) were solved in the non-shared L1 pairs. This shows that the test-takers in the shared L1 pairs were always capable of solving the communication problems that occurred, and they seemed to succeed in solving them, while the test-takers in the non-shared L1 pairs seemed not to be always successful.

The inter-coder agreement on non-engagement for the shared L1 pairs was .80, and that for the non-shared L1 pairs was .83. The inter-coder agreements on miscommunication for the shared and non-shared L1 pairs were .74 and .97, respectively. For a number of solved communication problems on non-engagement, the inter-coder agreements for the shared and non-shared L1 pairs were .75 and .88, respectively. The inter-coder agreement for the solved communication problems regarding miscommunication for the shared L1 pairs was .78 and for the non-shared L1 pairs was .90. Although the agreement rates ranged from .74 to .97, the agreement rates were generally acceptable.

A non-parametric Wilcoxon Signed Rank Test was utilised to investigate the statistical difference in the number of two types of communication problem: non-engagement and miscommunication between the shared and non-shared L1 pairs.

Table 5.8: Statistical differences for communication problems, non-engagement and miscommunication of the test-takers in the shared and non-shared L1 pairs

Communication problems	Type	Mean	Med	SD	Min	Max	Wilcoxon
Non-engagement	**SL**	.20	0	.41	.00	1.00	Z = -1.56
	NSL	.50	0	.76	.00	2.00	*p* = .12
Miscommunication	**SL**	.70	.50	.86	.00	3.00	Z = -1.63
	NSL	1.45	1.00	1.43	.00	5.00	*p* =.10

Note: SL refers to shared L1 pairs.
NSL refers to non-shared L1 pairs.

As shown in Table 5.8, there were no differences in the number of communication problems encountered between the test-takers in the shared and non-shared L1

pairs. However, the notable difference between the two groups is that while the shared L1 pairs managed to solve all problems, the ratios of solved communication problems for the non-shared L1 pairs were 70% for the non-engagement problems and only 40.48% for the problems involving misunderstanding.

Some examples of the two types of communication problem and how the test-takers in the non-shared L1 pairs responded to those problems are now presented in the following section.

5.1.2.1.1 Non-engagement

One of the communication problems that the test-takers in both shared and non-shared L1 pairs encountered was non-engagement, when one of the test-takers did not respond to their partner's question. This often resulted in their partner attempting to utilise various strategies to encourage them to deliver their idea by repeating or paraphrasing the particular question, repeating the keyword (see details in Table 5.7). Excerpt 15 is an example of non-engagement in the non-shared L1 pairs. P03 and T03 were talking about professions. P03 started asking T03 about this using lengthy questions which might have confused T03. As presented in Line 09, T03 did not answer P03's questions. He was quiet for a while and then responded with "*yeah*" without delivering his answer. Therefore, P03 paraphrased his answer, and this made T03 understand P03's question and be able to answer the question.

Excerpt 15

Topic: Professions (P03=Pakistani male 03, T03=Thai male 03)

	L01	P03:	what- >what do you think yourself which professions< is the
	L02		most difficult to achieve (.) and .hh ah: (0.3) >what do you
	L03		think is the most difficult< that level of level of high? .hh so
	L04		what do you think of different professions here? ((pointing
	L05		at pictures)) (.) so after that >I'll let you know about my my
	L06		opinion< for- for for in my- my opinion what what professions
	L07		to be difficult for me. (0.3) so (.) ((looking at a partner)) what
	L08		what do you think?
→	L09	T03:	(0.5) yeah
	L10	P03:	ah: if you look at the first picture, what wha- ah what (.)
	L11		profession is more difficult? ((looking at a partner))
	L12	T03:	.hhh ah:: I think this side ((pointing at a picture)) (0.9) this
	L13		professional will (0.9) er hhh it's not (0.5) not (.) ah develop
	L14		for the future

5.1.2.1.2 Miscommunication

According to Gass and Varonis (1991), miscommunication can be categorised into two subcategories: (1) misunderstanding and (2) incomplete understanding. However, this study only investigated miscommunication in broad terms. That is, it considered a mismatch between the intention of a speaker and the interpretation of a listener either fully or partly. The test-takers in both types of pair employed various strategies to solve their miscommunication during conversing, for instance, making a clarification request or confirmation check, giving an explicit response, using a paraphrase, and keyword repetition, as shown in Table 5.7. An example of miscommunication which occurred in the shared L1 pair (P05, P06) is shown in Excerpt 16. The Urdu L1 test-takers (P05, P06) were talking about professions. P06 was trying to explain that being a scientist was the most difficult profession to get to the top of, but he was including several pauses (Lines 29–31). He compared it with being a doctor and a singer, and extended his idea on being a singer. P05 summarised what P06 had said and requested confirmation from P06 (Lines 34–35). P06 confirmed P05's understanding by saying "*yeah*" (Line 36). When P05 continued extending his idea on being a singer (Lines 37 and 39), P06 realised that P05 misunderstood him. Hence, he repaired by saying "*no no no*" and clarified his idea (Line 40).

Excerpt 16

Topic: Professions (P05=Pakistani male 05, P06=Pakistani male 06)

L29	P06:	i think so (0.3) this ((pointing at a picture)) the scientist (.)
L30		is much much difficult profession (0.3) that (.) to the top
L31		(0.3) i think so. (0.9) It's much difficult than the doctor (0.3)
L32		than singer (.) singing is (.) most popular in most er (.) most
L33		(.) part of the world but (0.5) scientist
L34	P05:	ok you think that singer singer er singing (0.3) ((pointing
L35		at a picture)) is the most popular in the world
L36	P06:	Yeah
L37	P05:	technique there is a lot of competition
L38	P06:	y(h)ah ha ha
L39	P05:	and she need to be work hard-
→ L40	P06:	no no no (0.5) i think scientists ((pointing at a picture))
L41		er:: have to be work hard because it it is er: totally mentally
L42		job (0.4) we have to be mentally presence (.) they don't er:
L43		celebrate and holidays

L44 P05: yes of course

In the stimulated recall interview with P06 about what he was thinking when he said "*no no no*", he reported, "*I thought he understood what I said that being a singer was the most difficult profession to get on the top. But when he talked more, I knew he understood that I thought being a singer was the most difficult profession. That's why I said 'no no no' and explained more about it."*

This example illustrates that the communication problem that occurred did not come from P05's difficulties in comprehending P06's utterance, but resulted from his misinterpretation of P06's message because the way P06 delivered his idea was rather ambiguous. P06 started saying that being a scientist was the most difficult profession to get to the top of, compared with being a doctor or a singer (Lines 29–32). Then he talked more about a singer being the most popular in the world (Lines 32–33). This deviation might have confused P05. However, P6 managed to get the topic back on track in Line 40. This example obviously shows that when the communication problem occurred in the shared L1 pairs, they managed to identify the source of the problem easily and reached a mutual understanding very smoothly.

Such proactive attempts to understand a partner when the communication problem occurred did not seem to happen in the non-shared L1 pairs as frequently as in the shared L1 pairs. One of the ways in which the non-shared L1 pairs behaved when faced with communication breakdowns or failures to comprehend their partners was being quiet with or without gesturing and back-channelling. Excerpt 17 shows an example which clearly illustrates that a test-taker in a non-shared L1 pair did not answer a question because she did not comprehend her partner's question; she just kept quiet and looked at her partner, as shown in Line 64. This lengthy silence without answering a question signalled that P14 did not understand T14's question, which led T14 to repeat the words "*close friends*", since she thought that her partner did not understand these key words. After that, P14 understood the question and was able to answer it.

Excerpt 17

Topic: Friends (T14=Thai female 14, P14=Pakistani female 14)

L61 T14: yes ((nodding head)) and (2.5) .hhh heh heh heh (1.0)

L62		((looking at pictures)) er: do you have close friend /kloʊs flen/?
L63		((looking at partner))
→ L64	P14:	(1.3) ((looking at a partner))
L65	T14:	close friends
L66	P14:	yes (.) i have close friend (0.3) but a few here but more
L67		than er:: back home and mm:: .hh you know here
L68		everyone is so busy with their job and study. they don't
L69		(0.3) get er:: enough time to (0.4) to be with you (0.3) so
L70		we have less of time here for friends but (.) er: back
L71		home i have many friends but . hhh i cannot call them all
L72		the time (.) as the time's different and (.) er: i'm
L73		here and we cannot discuss usually as we were there

In the stimulated recall interview with P14, she reported that at first she did not understand that T14 was asking her about a "*close friend*" so she was quiet. P14 might have found it difficult to understand T14's accent when she first heard it due to P14's unfamiliarity with the accent of her partner, who was from a different L1 background (Harding, 2012), though she reported in the questionnaire that she had a neutral opinion on her familiarity with English spoken by Thai L1 speakers. T14's pronunciation of "*close friend*" indeed had some L1 influence as it was pronounced as "/kloʊs flen/". Regarding P14's listening proficiency, she got 20 out of 37 points in the listening test and was in Band 5.5 in the listening part of the IELTS. Her listening proficiency was therefore not great, but it was not likely that she had difficulty in understanding such a frequently used phrase. Indeed, P14 knew the meaning of "*close friend*" and understood it when she heard it the second. This result may be explained by Jenkins's (2002) finding that NNS–NNS interaction with a speaker with a below-bilingual proficiency level possibly fails to use contextual cues to solve difficulty in listening comprehension which derives from the pronunciation errors of their partner. This evidence echoes one of Field's (2004) findings on listening, which is that inefficient NNS listeners interpret the meaning of words or sentences from what they hear and do not utilise the context to help them understand the speech message. In accordance with this finding, P14 could not answer T14's question because the key words "*close friend*" were mispronounced as "/kloʊs flen/". Therefore, she was quiet and did not respond when she first heard these words. Pronouncing a word with an "*r*" in it seems to be problematic for Thais (Phootirat, 2012), and this possibly causes communication

problems between Thai and non-shared L1 interlocutors who are not proficient in English.

As discussed in Section 5.1.1.3, back-channelling is one sub-category of interactive listening, which is a part of successful speaking and listening interaction. Back-channelling can be used to encourage the speaker to continue speaking, and Ducasse and Brown (2009) named this category of listening "*a supportive listening*", as defined at the beginning of this chapter. However, Ducasse and Brown also note that back-channelling does not necessarily mean that a listener really understands their partner. This was also found to be true in the current study.

An example of non-shared L1 pairs' use of back-channelling merely for supportive listening without completely understanding their partner's speech is a conversation between T14 and P14 in Excerpt 18. They were talking about friends. T14 was trying to explain the advantages of having friends. However, she seemed to have difficulty in delivering her idea, as observed from her frequent pauses during her turn and her hand gestures when she could not express her ideas or think of appropriate vocabulary. P14 might have noticed that T14 had difficulty in explaining her idea, but P14 did not help her. P14 instead used back-channelling to encourage T14 to keep speaking. It is not likely that P14 comprehended T14 completely as T14's utterance even included the non-English word "*fortuner*" (Line 12). Nevertheless, P14 did not attempt to reach mutual understanding, and P14 instead initiated her own topic when T14 gave her a turn.

Excerpt 18

Topic: Friends (P14=Pakistani female 14, T14=Thai female 14)

L07	T14:	mm:: for me my friends (1.0) make me get better
L08		always make me get better ((waving hands))
L09	P14:	Ok
L10	T14:	when i have some problems
L11	P14:	uh huh
→ L12	T14:	(0.5) and (0.8) .tch! when i live here (.) i have fortuner
L13	P14:	uh huh
L14	T14:	(0.5) i (.) i can ((moving hand)) (0.3) practice my
L15		Speaking
L16	P14:	uh huh
L17	T14:	and (.) what about you? ((pointing at a partner))
L18	P14:	for me my friend is when i'm sad (0.3) i can call

In the stimulated recall interview with P14, she said, "*I was listening to my partner and thinking of how to formulate my idea in the same time.*" This shows that P14 did not pay much attention to her partner and did not show her interactive communication apart from superficial back-channelling because she was worrying about how to present her idea. This might be the cause of her giving no assistance to her partner by supplying a word or requesting clarification when her partner pronounced an unintelligible word.

This type of interactive listening, superficial supportive listening, was also found in the non-shared L1 pair (Thai L1- and Gujarati L1-speaking test-takers) from the pilot study (see Section 3.3.5.3.2). Back-channelling was used by a Thai L1-speaking test-taker while listening to her Gujarati L1-speaking partner's lengthy talk about Formula One car racing. She admitted that she did not understand her partner, but she pretended to comprehend him by using back-channelling and sometimes nodding her head.

To conclude, when a communication problems occurred, it seemed that the test-takers in the shared L1 pairs solved all problems, while the test-takers in the non-shared L1 pairs did not always succeed in doing so.

5.1.2.2 Understanding unclear utterances and incorrect word use

One of the differences in communication patterns related to interactive listening between the shared and non-shared L1 pairs was understanding unclear utterances and incorrect word use. In the shared L1 pairs, even unclear utterances and incorrect word use were understood by a partner, but this did not happen in the non-shared L1 pairs. In addition to being caused by test-takers' limited listening proficiency, misunderstanding and communication breakdown can also occur as a result of unclear utterances, incorrect word use and incorrect use of grammar. It is interesting that these factors did not seem to be very problematic for the test-takers in the shared L1 pairs. They seemed to understand their partner's utterances easily, even when they contained errors, as shown in Excerpt 19, which is a conversation between Thai L1 test-takers.

In Line 69, T07 asked T08, "*how did your close-close friends*?", mistakenly omitting a main verb. Omitting a main verb is not considered to be a Thai L1

transfer, as the Thai language does not allow verb omission. Without a main verb in a sentence, a message is unintelligible in Thai and English. Nevertheless, T08 could answer the question correctly. As a result, the conversation continued successfully.

Excerpt 19

Topic: Friends (T07=Thai male 07, T08=Thai male 08)

L66	T08:	for me i know (.) i know many people ((moving
L67		hands)) but- but (.) ah:: the best friends of mine only
L68		a:bout two or three people only i can talk to (0.3)ah[::
→ L69	T07:	[how
L70		did your close-close friends?
L71	T08:	i study together about the (0.3) more than ten years

When T08 was interviewed, he reported that he did not even notice that the question was grammatically incorrect. He somehow interpreted correctly the partner's intention to say "*[H]ow did you meet your close friends?*"

Another example is given below in Excerpt 20, which shows a part of a conversation between a Thai L1 pair (T19, T20) about friends. In Line 33, T19 intended to say, "*[A] girlfriend and a boyfriend do something (activities) together*", but she mistakenly used the verb "*make*" instead of the verb "*do*". Nonetheless, T20, without even making a clarification request, understood what T19 intended to say and then T20 delivered her idea about it. In Thai, "*make*" and "*do*" have the same meaning and it is a common mistake of Thai speakers of English to use these two verbs interchangeably.

Excerpt 20

Topic: Friends (T19=Thai female 19, T20=Thai female 20)

L28	T20:	that's it and this yeah it's like (0.9) [playing music=
L29	T19:	[er:::
L30	T20:	=[together
L31	T19:	[what do you think about (0.7) er::: (0.3) girlfriend and
L32		(.) boyfriend ah::: ((moving indexes closely)) (0.5)
→ L33		make something together?
L34	T20:	i think for the girl they is like maybe share feeling
L35	T19:	mm:::
L36	T20:	or talk something because like girl more talkative .hh than
L37		than [guy so they will more share .hh share feeling than =
L38	T19:	[huh huh huh

L39	T20:	= than guy .hhh but for guy they just (.) ok go together
L40		or play the music but they not [talk yeah talk too much
L41	T19:	[not not mm:: ((nodding
L42		head))

In T20's stimulated recall interview, T20 reported, "*I knew what she meant. In my language, 'make' and 'do' have the same meaning*." This showed that T20 could understand T19 easily because they have the same L1 background knowledge and she knew that it was a common mistake in the use of English made by Thai L1 speakers. May (2007) also reported that test-takers from the same L1 background understood each other easily while speaking in an L2 when raters had difficulty in comprehending their talk.

By contrast, there was no such instance in non-shared L1 pairs. Test-takers in non-shared L1 pairs could not comprehend unclear utterances or answer unclear questions without explicit negotiation of meaning, as exemplified in Excerpt 21. In this excerpt, T05 asked P05 a question about the advantages of having friends, and he also asked, "*[I]s there anything normal?*" (Line 2), which was ambiguous and did not relate to the previous question. P05 was quiet for a while and said "*Pardon?*" while looking at her partner to signal that he did not understand. This is an explicit indication of P05's non-understanding of a message which T05 intended to deliver. Therefore, T05 clarified what he meant in Line 6.

Excerpt 21

Topic: Friends (P05=Pakistani male 05, T05=Thai male 05)

L01	T05:	let's start with the first question that that what are-
L02		what are the [advantages of having friends? =
L03	P05:	[advantages
L04	T05:	= is there anything (.) normal?
→ L05	P05:	(0.7) pardon? ((looking at a partner))
L06	T05:	it's easy yeah ((moving hands)) (.) advantages of
L06		friends, friend is (0.3) i think it's a good thing, ◦isn't
L07		it? ((scratching head))
L07	P05:	friends in our life is gonna be cheerful (0.3) having
L08		[joyful
L09	T05:	[share (0.3) experience together ((moving hands))

In the stimulated recall interview with P05, he reported, "*I said pardon*

because I didn't get his question. Luckily he explained it, so I could answer it." His listening proficiency was relatively high; he scored 26 out of 37 in the listening test, and he was in IELTS listening Band 6.5. He was unable to answer the question, not because of his listening proficiency, but because the question was not related to the prior utterance of the partner. He was confused by the question, which was unclear and unrelated to the context. He could not guess the speaker's intention either.

From the above examples, it might be suggested that even unclear utterances could be understood easily in the shared L1 pairs without relying on explicit meaning negotiation. This could be because the shared L1 pairs may be aware of English common mistakes made by the same L1 speakers. They may utilise the same L1 linguistic background (Bent and Bradlow, 2003; Hahn and Watts, 2011; Kachi, 2004) and cultural background (Kachi, 2004) to interpret their partner's speech. Furthermore, the questionnaire result of this study disclosed that the test-takers were more familiar with English spoken by the shared L1 speakers than by the non-shared L1 speakers (see Section 3.4.1.1 and Appendix 10), which seemed to support the findings here. Familiarity with the English spoken by shared L1 speakers seemed to enhance mutual understanding even when an utterance was unclear or ambiguous. However, this did not guarantee that the rater would fully understand the test-takers (e.g. May, 2009), since the raters who participated in this study were not familiar with English spoken by Urdu or Thai L1s (see Section 3.4.1.2).

5.1.2.3 Misunderstanding because of having different cultural background

The last pattern, which is categorised by the differences in communication patterns related to interactive listening between the shared and non-shared L1 pairs, is misunderstanding because of different cultural background. It was observed only in the non-shared L1 pairs. The test-takers in the non-shared L1 pairs seemed to have difficulty in understanding their non-shared L1 partner because of their different cultural backgrounds. It could even cause a communication problem. For instance, Excerpt 22 illustrates miscommunication between P10 and T10 about how to make new friends. T10 suggested that his partner, P10, should arrange a party. P10 imagined that alcohol would have to offer at such a party, which is prohibited

according to his religious beliefs. From the stimulated recall interview with P10, he disclosed, "*I am Muslim and our religious don't allow us drinking alcohol. I tried to tell my partner about it.*" However, P10 was trying to explain his reasons only very implicitly, referring to his lack of skill in arranging parties and his lack of self-confidence. T10 seemed not to understand his partner's hidden problems with arranging the party and proposed that P10 should move to a new topic (Line 69).

Excerpt 22

Topic: Friends (P10=Pakistani male 10, T10=Thai male 10)

L45	P10:	=and er: (.) ha ha i have a .hh very short list of friends
L46		(.) yeah i only have two or three friends ha ha [and the=
L47	T10:	[uh huh
L48	P10:	=mm:: (.) from my [part
L49	T10:	[you- you can do the party a lot man
L50		[yeah if you want to make a lot of friends [ha ha ha ha
L51	P10:	[ha ha ha [yeah ha ha ha
L52		actually problem's that i'm not good at party [ha ha ha
L53	T10:	[oh yes
L54	P10:	er:: i have not having some (0.3) lots of the friends
L55		((moving hands)) [and (.) but er:: (.) in my start (0.3)
L56	T10:	[mm::
L57	P10:	[like ah:: i am here as a new here [er:: one month ago in=
L58	T10:	[uh huh [uh huh
L59	P10:	the uk er::: and er:: i also (0.3) er: get some (.) .hh kind of
L60		(.) the mm::: (.) lack of confident [also having some=
L61	T10:	[uh huh uh huh
L62		((nodding head))
L63	P10:	=like (0.3) to er:: is having problem ((moving hands)) to
L64		make with er: (.) the friends [ev:ery friend and what's=
L65	T10:	[uh huh
L66	P10:	=happening
L67	T10:	right uh huh
L68	P10:	and er::: [.hh i must
→L69	T10:	[so let's go to the next [hih hih hih

The misunderstanding occurred because P10 did not explicitly mention the real reason that P10 had for being reluctant to arrange the party, which was related to his cultural background. This indicates that explicit explanation is the key to successful communication in the non-shared L1 pairs who had different cultural backgrounds.

5.2 Additional interactional features between test-takers in shared and non-shared L1 pairs

The researcher observed two additional interactional features that relate to test-takers' L1 backgrounds: (1) the test-takers provided their L1 back-channelling while listening and (2) inserted L1 words into their speech.

5.2.1 Providing their L1 back-channelling while listening

When the test-takers were listening to their partner, back-channelling was utilised to show their engagement with and comprehension of what their partner had been saying. Sometimes they uttered their L1 back-channelling automatically. This was evidenced only in Thai L1 test-takers. Thai back-channellings were found three times from two Thai L1 test-takers. The interesting point is that L1 back-channelling occurred only when those Thai L1 test-takers were paired with a non-shared L1 partner. Below is an example of Thai back-channellings which was pronounced unconsciously by a Thai test-taker.

P08 and T08 were talking about professions. P08 was expressing his idea of how to be successful in an artistic profession. While listening to P08, T08 presented his interactive listening by supplying the words as shown in Line 29 and pronouncing Thai back-channelling (/ɔ́ː/), as shown in Line 34, which has a similar meaning to "*I see*" in English.

Excerpt 23

Topic: Professions (P08=Pakistani male 08, T08=Thai male 08)

L26	P08:	artist er::: to become successful artist eh: it depends
L27		on (0.3) basically depends on .hhh er:: practice
L28		[the- the er:: how much the percent ex- er:: =
L29	T08:	[practice
L30	P08:	exam ((moving hand)) or .hhh (.) give them more time
L31		to practice then er: you(.) you er: get familiar of the
L32		work or exam (.) practice makes something perfect
L33		[the same thing is ah:: very applicable in the case=
→ L34	T08:	[/ɔ́ː/
L35	P08:	=of artist

In the stimulated recall interview with T08 about why he pronounced Thai back-channelling, he reported, "*I didn't know why I said it. I didn't realise that I*

pronounced Thai word. I intended to listen to my friend and tried to understand him." T08 unconsciously uttered Thai back-channelling perhaps because he was putting a lot of effort into comprehending his partner's speech. Therefore, he was not aware of his L1 back-channelling use, which he pronounced unconsciously. In addition, the L1 back-channelling use seemed not to cause any communicative problem for his partner, who did not share L1 background with him.

5.2.2 Inserting an L1 word while speaking

Not only did Thai L1 test-takers utter L1 back-channelling, but one Thai test-taker also unconsciously inserted an L1 word. This was observed only in one Thai L1 test-taker (T19), as illustrated in Excerpt 24. It is a conversation between P19 and T19. They were talking about professions. T19 seemed to have difficulty in explaining her idea, as shown by her many pauses and hesitation utterances, e.g., "*er*" and "*ah*", during her speech, while P19 patiently listened to T19 until P19 heard T19 say the word "*professor*", which did not seem to be relevant to a business context. This led P19 to ask for clarification from T19. In responding to the request, T19 unconsciously pronounced the Thai word "/*bæb/*" (Line 79) which means "*to be like*" in English.

Excerpt 24

Topic: Professions (P19=Pakistani female 19, T19=Thai female 19)

L66	P19:	[yah I can't say anything about it because
L67		she's doing sport sciences [so maybe it's easy for her
L68	T19:	[yeah yeah if- if someone
L69		ah::: (0.5) ah:: (0.6) get training i think (0.4) get
L70		training a lot ah::: .hh (0.7) maybe everyone can (0.9)
L71		((moving hands)) play or (0.4) can do anything about
L72		exercise or sport (0.3) ah:: it's well (.) and: but for me
L73		i think a businessman ((pointing at a picture)) is .hh
L74		(0.4) ah:: (0.7) a few of people (.) ah: can (1.2) make
L75		((moving hand)) (.) them to profess- professor (.) er::
L76		and (0.3) .tch! (0.6) [er:::
L77	P19:	[what do you mean by
L78		businessman is professor?
→ L79	T19:	er::: (.) like /bæb/ er::: ((looking at a ceiling)) (1.5)
L80		professional ((looking at a wall)) (0.7) like
L81		((waving hands)) (0.7)
L82	P19:	Professional

L83 T19: yeah yeah professional

When T19 was interviewed after the conversation between her and P19 had finished, she said, "*I couldn't explain my idea. I was thinking of Thai words before translating them into English. It was very difficult for me to speak. But I was able to understand my partner. She talked a lot.*" Therefore, the Thai word seemed to be unintentionally pronounced in the mental translation process, probably because she felt more relaxed and comfortable than when interacting with a non-shared L1 partner. T19 possibly had difficulty with the demands of the cognitive process used to decode a spoken message and compose her own speech (Field, 2011) which resulted in her unintentionally producing her L1 vocabulary. Furthermore, the interaction between this pair was noted by raters to have an asymmetric interaction pattern (dominant vs passive), showing low equality in interaction (Galaczi, 2004). T19, who allowed her partner to dominate the speaking floor, was a passive speaker, and P19, who was more talkative than T19 and rarely allowed T19 to speak, was a dominant speaker. This was possibly because T19 had limited linguistic ability (IELTS speaking Band 5.5 and listening Band 5, and 17 out of 37 points in the listening test) and felt uncomfortable interacting with her non-shared L1 partner (based on her response to the questionnaire, she disagreed with the familiarity of the English spoken by Urdu L1 speakers).

While no general conclusion can be drawn from the use of only three test-takers' use of their L1, it is interesting to find such examples of L1 use in the non-shared L1 pairs. Unlike Jenkins's (1997) study, in which shared L1 pairs adjust their pronunciation by producing a more L1 accent to increase mutual intelligibility, this study observed the incidences of L1 use only in the non-shared L1 pairs. The two excerpts above that showed L1 back-channelling and the use of an L1 word in the non-shared L1 pairs seem to indicate that such unconscious usage relates to these Thai test-takers' limited English proficiency, their limited use of English in their daily lives and their stress in the testing context. The test-taker in Excerpt 23 used L1 back-channelling when he was trying very hard to comprehend his partner's speech, who spoke English with a different accent from him, and the test-taker in Excerpt 24 used an L1 word by mistake when she had difficulty expressing her idea in English. In contrast, even if the Urdu L1 test-takers in this study had a similar

level of English proficiency, they did not use their L1.

5.3 Raters' perceptions of shared and non-shared L1 pairs' interaction and their listening proficiency

This section reports the raters' perception gained from stimulated recall interview of the shared and non-shared L1 pairs. As described in Section 3.4.4, the main themes emerged from the data can be classified as follows:

1 The shared L1 pairs seemed more relaxed and more interactive than the non-shared L1 pairs;
2 The shared L1 pairs increased mutual understanding and attempted to solve communication problems; and
3 The shared L1 pairs showed additional features of achieving successful interactive communication.

5.3.1 Shared L1 pairs were more relaxed and interactive than non-shared L1 pairs

The raters noticed that the test-takers in the shared L1 pairs seemed to be more relaxed and more interactive than those in the non-shared L1 pairs. The raters' comments recurrently referred to how comfortable test-takers appeared when talking to their partners and how interactive their interaction was. Table 5.9 below lists some of the comments on this issue. Both raters repeatedly reported that the non-shared L1 pairs seemed to be less interactive than the shared L1 pairs.

Table 5.9: Raters' stimulated recall interview results regarding test-takers' relaxation and interactivity when interacting with shared and non-shared L1 partners – selected comments

Shared L1 pair	Non-shared L1 pair
• *"Test-takers seem to feel more comfortable when interacting with a shared-L1 partner"* (Rater 1); • P03 and P04 pair: *"P03 uses a variety of language to recall something that happened, but there are many grammatical errors. P04 asks and answers questions and sometimes he picks up on his partner's word and develops it, which elaborates the flow of the conversation and helps it to be like a genuine conversation"* (Rater 1); • P19 and P20 pair: "*P19 picks up on words to continue and engage in the conversation and have lively interactive discussions, starts off the conversation, then has to interrupt a partner to put point of view. Sometimes both speak together. Reacts to a partner's opinions. P20 repeats what her partner said, picks up on her partner's opinion, and extends them with her own ideas, [she] almost dominates the discussion rather than giving her partner much time to speak as much as she would like. So no questions and answers are necessary as they sometimes speak at the same time. Picking up on a partner's opinions is an enthusiastic interaction*" (Rater 1); and	• P03 and T03 pair: "*P03 speaks very fast and provides little interaction. He initiates and responds to his partner's idea, though he sometimes dominates the conversation*" (Rater 2); • P03 and T03 pair: "*T03's responses are very short. He listens rather than speaks. He nods and shows his agreement, but does not respond orally very much. He lacks interaction which does not extend the conversation. He could possibly be pretending to understand"* (Rater 1); • P04 and T04 pair: "*P04 does not really hold a discussion. He asks questions, makes no comments, only responds 'yes' and 'no' to a partner. He just keeps speaking. He doesn't encourage his partner to speak*" (Rater 1); • P04 and T04 pair: "*T04 uses a great variety of language and expands his ideas. He just says what he thinks, but there isn't much interaction*" (Rater 1); • P04 and T04 pair: "*It's difficult to tell who is better at interaction. There is not much interaction. One uses simple language and the other one uses more complex language*" (Rater 1); • P19 and T19 pair: "*P19 dominates the interaction and talks over her partner. She dominates the conversation and doesn't allow her partner to talk. There is almost no interaction. She doesn't respect turn taking rights of her partner at all. She obviously loves like the sound of her own voice*" (Rater 2);

Shared L1 pair	Non-shared L1 pair
• T03 and T04 pair: "*They tend to encourage each other's talk via smiling and back-channelling use, for example, 'mm', 'huhuh', 'yeah'. They take turns to talk. No interruption. T03 shows more of an interest in the speech of T04 by smiling, saying 'huhuh', 'yeah' and 'it's true'. T4 can initiate, respond, maintain and develop the conversation with no support from his partner, but sometimes he doesn't link his ideas to his partner's"* (Rater 2).	• P19 and T19 pair: "*T19 takes no part in the discussion. Few contributions and comments. No questions. She uses a lot of hesitations 'er', as an indication that she is searching for words. She seems to understand her partner, but she lacks confidence to speak English which leads to the inability to engage in a more interactive communication*" (Rater 1); • P19 and T19 pair: "*It's not really an interaction. One (P19) dominates the conversation while the other one (T19) rarely speaks, and there is no conclusion at the end*" (Rater 1); • P19 and T19 pair: "*T19 looks more serious, less relaxed and more attentive than when she talked to her Thai interlocutor. She doesn't try to initiate or present her ideas. She allows her partner to dominate the conversation. P19 obviously dominates the interaction and talks over her partner. She doesn't allow her to talk. She doesn't even ask for any opinions from her interlocutor. It isn't an interaction*" (Rater 2); and • P20 and T20 pair: "*T20 looks more serious, less relaxed and more attentive than when she talked with her Thai interlocutor. This may be because she feels more comfortable talking to a partner from the same L1 background than from the different L1 background. T20's speaking role when talking with the Pakistani interlocutor is less than when talking with Thai interlocutor*." (Rater 2).

As shown in the above data, both raters perceived that the test-takers in the shared L1 pairs were more comfortable interacting with each other and produced more interactive talk than the non-shared L1 pairs. As observed from the test-takers' non-engagement in their communication (see Table 5.7), all four non-engagement behaviours of the test-takers in the shared L1 pairs occurred because they were searching for ideas, while the causes of non-engagement behaviours of test-takers in the non-shared L1 pairs were because of their limited listening proficiency (5 out of 10 non-engagements) and because they were searching for ideas (5 out of 10 non-engagements). Three test-takers in the non-shared L1 pairs did not extend the partner's ideas but only initiated their own topics. Some test-takers, especially Thai L1 test-takers, allowed their partner to dominate the conversation while listening to them with or without providing back-channelling or nodding their head.

5.3.2 Shared L1 pairs' increased mutual understanding and their attempts to solve communication problems

Both raters reported that the test-takers in the shared L1 pairs seemed to understand each other easily. They tended to encourage each other to talk through interactive listening devices (i.e., back-channelling and supplying vocabulary), and when they faced difficulty in delivering ideas, their partner helped to extend the ideas, which made the conversation continue. On the other hand, the test-takers in the non-shared L1 pairs did not always try to solve communication problems, and this was also captured in the raters' stimulated recall interview data.

Table 5.10: Raters' stimulated recall interview results on test-takers' mutual understanding and attempt to solve communication problems when interacting with shared and non-shared L1 partners – selected comments

Shared L1 pair	Non-shared L1 pair
• P03 and P04 pair: "*P03 uses a variety of language to recall something that happened, but there are many grammatical errors*" (Rater 1); • T19 and T20 pair: "*T19 understands what her partner said, but it seems that she has limited command of English. She has a linguistic problem*" (Rater 1); • T03 and T04: "*T04 does not put stress on the ending sound of the final consonant, for example, 'l', 'r', 'th', which makes it difficult to understand*" (Rater 1); • T19 and T20 pair: "*T20's pronunciation makes her speech difficult to understand. Her projection of English is good. She can maintain the interaction, deliver her opinions, support and respond to her partner's idea with jerky delivery*" (Rater 1); and • T19 and T20 pair: "*The one [T20] who is more talkative tends to help the other one (T19). When she is confronted with speaking trouble, she expresses her difficulty by admitting it and asking for help"* (Rater 2).	• P03 and T03 pair: "*Sometimes, P03 tries to clarify his question, but fails to do so because he asks his partner with a complicated question*" (Rater 2); and • P19 and T19 pair: "*T19 has no room to talk and doesn't show any signs of wanting to talk at all. She just listens to her partner and waits for her turn. No interruption ever. No argument given. T19 always smiles and listens to her partner. She doesn't try to interrupt or take a turn from P19. There are a lot of pronunciation problems, for example, final consonant sounds. One ([P19] often has extra final –s; the other (T19) leaves it in some words. 'Sport science' which sounds like 'spot scient /spɒt scaɪns/' will never be understood by any other English speakers. But the interlocutor (P19) doesn't show any sign of misunderstanding because she only initiates her idea without extending to her partner's idea*" (Rater 2).

The above comments from the raters indicate the test-takers in the shared L1 pairs understood each other easily. They were collaborative in co-constructing the conversation. In addition, they helped each other to develop the conversation by asking questions, responding to and extending their partner's ideas. Besides this, they assisted each other to solve communication problems and maintain the conversation. For example, in the case of a silence or when their partners awkwardly responded to a question, they assisted by clarifying the question (see Excerpt 14), while in the non-shared L1 pairs the test-takers did not talk until one test-taker could not tolerate the silence any longer and attempted to break the silence by repeating vocabulary or presenting his/her own idea (see Excerpts 19 and 20).

Additional data from raters' stimulated recall interviews adds further insights from raters into features of test-takers' interactive communication. Details of the additional features are portrayed in the following section.

5.3.3 Additional features of successful interactive communication

Features of the test-takers' interactional performance, based on the raters' perspective, which could not be included in the above two categories, were pronunciation, eye contact, fluency, interactive listening, using intelligible words and confidence. These features were considered to affect test-takers' interactional communication scores.

5.3.3.1 Pronunciation

Pronunciation was raised as a concern by Rater 1 as it might affect the comprehension of the test-takers' partner. Three pronunciation issues which made speaking difficult to understand were intonation, individual sound pronunciation and individual word stress. Furthermore, compulsive speaking also affected their listening comprehension.

- P03 and T03 pair: "*This one [T03] has a much smoother delivery than his partner [P03]. His utterance is like an English accent, because he speaks smoothly and his intonation is much smoother than this one [P03] who speaks jerkily, which affects his partner's understanding. He has a problem pronouncing 'the', 'l' and stressed the wrong syllable. These three issues of pronunciation make it difficult to understand: (1) intonation, (2) pronunciation of individual sound, and (3) stress on an individual word*" (Rater 1).

This issue tends to cause more difficulty in a partner's understanding when test-takers were from different L1 backgrounds.

5.3.3.2 Eye contact

Rater 1 pointed out that eye contact was important in interaction. A lack of eye contact possibly influenced the effectiveness of communication. Rater 1 assumed that a test-taker lacked eye contact because he was searching for ideas to deliver.

- P03 and P04 pair: "*P04 speaks with very little eye contact, hardly looking at his partner's face. Maybe he is thinking what to say next*" (Rater 1).

5.3.3.3 Use of fillers

Both raters commented on some test-takers' frequent use of fillers such as "*er*" while searching for words or when having difficulty in presenting their ideas. This seemed to affect the effectiveness of their interaction.

- P03 and P04 pair: "*P04 uses a lot of 'er', which I think he is searching for words, so his language is not continuous and has more errors in his speech. It's not natural to say much 'er'*" (Rater 1); and
- P03 and P04 pair: "*P04 utters lots of 'er'. He might have difficulty in delivering his ideas, or he might be searching for words*" (Rater 2).

5.3.3.4 Interactive listening

Both raters reported that showing interactive listening during a conversation was a part of successful interactive communication. It illustrated their understanding and their attention to their partner's speech.

- P19 and P20 pair: "*P19's helping her partner by saying the word 'destroy' presents her interactive listening. It is a part of judging how they succeed in interactive communication. But in this pair, I awarded P19 a score only 3 out of 5 because she seems to dominate the conversation*" (Rater 1);
- P19 and P20 pair: "*P20 seemed to be struggling with finding an appropriate word to describe the effect of having bad friends, and P19 said that word. It showed her interaction while listening. But I awarded her a score of 2 out of 5 in the interactive communication because she really dominates the conversation and hardly lets her partner speak. She also frequently interrupts her partner*" (Rater 2); and

- P20 and T20 pair: "*P20 shows her understanding to her partner's speech by saying 'yes', 'yes, you're right and then explaining her idea. It is encouraging to show her partner that she's listening to her*" (Rater 1).

5.3.3.5 Intelligible word use

Producing intelligible words is a fundamental factor of communicative achievement. When unintelligible words are uttered, it might cause difficulty in understanding. During paired interaction, one test-taker pronounced a word which was unintelligible for the raters because it was not an English word. As a consequence, it was not understandable by a partner of that test-taker, who was from a different L1 background.

- P19 and T19: "*I think T19 pronounces Thai words sometimes, for example, /bæb/, and I don't understand it*." (Rater 1).

5.3.3.6 Confidence

Test-takers' confidence in speaking was also important for the raters in assessing the test-takers interactive communication. A lack of confidence affected the interactional effectiveness, and the test-takers who lacked confidence were awarded low scores in the interactive communication category.

- P19 and T19 pair: "*T19 seems not to be confident enough to present her idea*" (Rater 1).

5.4 Summary and discussion

This chapter has presented and discussed the results gained from CA, test-takers' and raters' stimulated recall interviews.

Some similarities and differences in communication patterns related to interactive listening between the test-takers in the shared and non-shared L1 pairs were identified. Three types of similarities in communication patterns were (1) supplying relevant vocabulary, (2) demonstrating comprehension, and (3) use of back-channelling. The differences in communication patterns consisted of (1) causing communication problems (2) understanding unclear utterances and incorrect word use, and (3) misunderstanding because of a difference in cultural backgrounds.

The stimulated recall interviews with the raters indicated that the shared L1 test-taker pairs were more relaxed and produced more interactive talk than in the non-shared L1 pairs. Also, the shared L1 test-takers seemed to understand each other easily and helped each other to solve the communication problems, while the non-shared L1 test-taker pairs did not always do so. They seemed to initiate their own topic rather than extending or developing their partner's idea. This is what Galaczi (2004) calls "*parallel interaction*". The raters also commented that the performance of some non-shared L1 pairs was not interaction. It was possible that they had difficulty in understanding their partner's talk because they were not familiar with a different accent and/or they had different cultural backgrounds. The finding from the questionnaire, which showed that the test-takers were more familiar with English spoken by the same L1 speakers than speakers with different L1 backgrounds (see Section 3.4.1.1 and Appendix 11), seemed to support the difficulty in understanding English spoken by speakers with a different L1 background. Understanding a non-shared L1 partner was more demanding, and this might have resulted in producing less interaction. In short, this study showed that English spoken by the shared L1 partners facilitated more interactive communication, which supports previous studies that demonstrated that the L1 variable affected types of interaction in pairs (Bent and Bradlow, 2003; Fayer and Krasinski, 1987; Jenkins, 1997, 2002; Kachi, 2004; May, 2007; Van Engen et al., 2010). In addition, the raters also disclosed additional features of interactive communication: pronunciation, eye contact, use of fillers, interactive listening, intelligible word use and confidence, which were considered to be important factors for interactional achievement in pairs.

The raters' perceptions of the shared and non-shared L1 pairs' interaction and their listening proficiency could explain the findings gained from the quantitative analyses. The test-takers' listening test scores correlated with their paired speaking test scores in grammar and vocabulary and discourse management in the non-shared L1 pairs. It is particularly interesting that the higher listening proficiency they had, the better they were able to present their own ideas with relevant organisation and wide range of linguistic resources. The interactional pattern that the non-shared L1 pairs used tended to be "*parallel*" rather than "*collaborative*" as they did not interact

with their partner much. They seemed to focus on initiating their own ideas rather than interacting with or extending their partner's ideas. This might be because they were less comfortable interacting with a non-shared L1 partner than with a shared L1 partner, as suggested in the raters' comments. Unfamiliarity with the English spoken by non-shared L1 speakers, as illustrated in the questionnaire result, might also have contributed to their attitude towards collaborative communication. This was also evident in Isaacs's (2013) study, in which asymmetric passive test-takers perceived that their partner' pronunciation was really problematic for them when communicating with different L1 partners. In addition, asymmetric dominant test-takers felt more frustrated when interacting with their passive partner who had a different L1.

It was surprising that the interactive communication scores of the test-takers in the shared and non-shared L1 pairs were not significantly different, although the test-takers interacted with their shared L1 partner much more than with their non-shared L1 partner (according to the raters' stimulated recall interview data). This might be because of their limited speaking proficiency, which caused them to find interaction difficult even though they seemed to understand their shared L1 partner well. Furthermore, the fact that there was a small number of participants (N=40) is possibly one of the causes of there being no statistical differences in their interactive communication scores.

To be successful in paired interaction, the test-takers' listening proficiency was important since they had to comprehend their partner sufficiently in order to respond to their partner appropriately and effectively. Conclusions and implications of this study and recommendations for further studies are discussed in Chapter 6.

CHAPTER 6 Discussion and Conclusion

The key findings from Chapters 4 and 5 are summarised in the first part of this chapter. This is then followed by the implications of the findings and the contributions of the study. Finally, the limitations of the study, directions for further study and final thoughts are provided.

6.1 Summary of the study

This study investigated the relationship between test-takers' L1, their listening proficiency and their performance in speaking in pairs. The participants were 40 pre-sessional English language programme students (20 males and 20 females) from two L2 backgrounds (Urdu and Thai). Test-takers were asked to complete a demographic questionnaire and take a listening test, a monologic speaking test and paired speaking tests. The speaking tests (which included a monologic task and paired tasks) were derived from the collaborative tasks of the FCE and equivalent in terms of topical and linguistic demands in order to provide meaningful comparison between the results of the different test formats. To avoid possible confounding variables related test-taker characteristics in the paired speaking tests, the test-takers were matched with a partner of the same gender who had similar English speaking and listening proficiency. They took the paired speaking test twice: one with a shared L1 partner and one with a non-shared L1 partner. After their paired speaking performance, test-takers were interviewed individually, employing the stimulated recall interview method, about their experience of taking the paired speaking test. Their speaking performances were video recorded and transcribed following the CA conventions (Atkinson and Heritage, 1984). Their stimulated recall interviews were audio recorded and transcribed orthographically.

Video recordings of the test-takers' speaking performances were scored by two trained native-speaker raters based on the FCE speaking criteria. The raters awarded the test-takers' analytical speaking scores in four categories: (1) grammar

and vocabulary, (2) discourse management, (3) pronunciation, and (4) interactive communication and also provided written comments. Furthermore, the raters were interviewed about their rating for the interactive communication shown in eight selected video recordings of the test-takers' paired performance.

The mixed-methods approach was employed to analyse and triangulate different kinds of data. The SPSS program was utilised to analyse the demographic questionnaire data, the listening scores and the analytical scores of the monologic speaking test and the paired speaking tests in order to answer Research Questions 1 and 2. Spearman correlations were used to explore the correlations between investigated variables (listening scores vs analytical scores of the monologic speaking test, and listening scores vs analytical scores of the paired speaking tests in the shared L1 pairs and the non-shared L1 pairs). Non-parametric Wilcoxon Signed Rank Tests were utilised to investigate the differences between variables (the analytical scores of the monologic speaking test vs the analytical scores of the paired speaking tests, and the analytical scores of the paired speaking tests in the shared L1 pairs vs the analytical scores of the paired speaking tests in the non-shared L1 pairs).

To answer Research Question 3, CA was carried out to explore the communication patterns in the paired speaking performances which related to the test-takers' interactive listening abilities and their L1s. The most important results are presented in the following section.

6.1.1 Summary of the key findings

This section presents the summary and synthesis of the key findings gained from quantitative and qualitative data analysis that was done to answer the research questions.

6.1.1.1 Research Question 1

To what extent is test-takers' performance in paired speaking tests in shared and non-shared L1 pairs affected by their listening proficiency?

Findings

There was no statistically significant correlation between the test-takers' listening scores and paired speaking scores for the shared L1 pairs, while there were statistically significant (but only moderate) correlations between the test-takers' listening scores and paired speaking scores in grammar and vocabulary and discourse management categories in the non-shared L1 pairs.

In the shared L1 pairs, the test-takers' listening proficiency did not affect their paired speaking scores in any categories. It suggests that high listening ability does not advantage test-takers when interacting with a shared L1 partner. That is, the shared L1 test-taker pairs tended to understand each other easily, because they were able to infer what their partner intended to say regardless of their listening proficiency. This intelligibility benefit seemed to be because of an advantage of sharing L1 linguistic and cultural background knowledge and familiarity with English spoken by shared L1 speakers, which enabled them to understand each other easily regardless of their proficiency levels (Bent and Bradlow, 2003; Fayer and Krasinski, 1987; Jenkins, 1997, 2002; Kachi, 2004; May, 2007; van Engen et al., 2010).

In the non-shared L1 pairs, the test-takers with a higher listening proficiency tended to receive a higher score in the grammar and vocabulary category. Since grammar and vocabulary are components which normally contribute to a significant proportion of the total score variance in skill-based tests (e.g., Geranpayeh, 2007; Shiotsu and Weir, 2007), the relationship between the test-takers' listening scores and paired speaking scores in grammar and vocabulary was expected to some extent. Of more interest to this study is the finding that the better their listening proficiency, the more effectively they tended to manage discourse in non-shared L1 pairs. This suggests that when the test-takers understood their partner's speech better, they could extend their partner's speech more coherently and initiated their own talk in a way that was more relevant to their partner's topic/idea.

Surprisingly, the test-takers' listening scores did not correlate with their paired speaking scores in interactive communication. This indicates that learners' listening ability does not necessarily help them display their proficiency in interactive communication skills. Although the test-takers were required to express

and exchange their opinions on the topics with their paired partner, the analysis of interactional data suggested that some pairs expressed their own ideas without extending their partner's ideas, which is identified as "*parallel interaction*" (Galaczi, 2004). They received low scores for interactive management by raters, regardless of the extent to which they understood the partner's utterances.

6.1.1.2 Research Question 2

Are there any differences in speaking scores when test-takers are paired with shared L1 partners as compared to (when they are paired with) non-shared L1 partners?

Findings

There was no statistically significant difference in any analytical categories between the two types of pairing.

No statistical differences were found in the paired speaking scores that test-takers received when they were paired with a shared L1 partner as compared to when they were paired with a non-shared L1 partner. Based on the statistical results, therefore, pairing with a shared L1 partner or a non-shared L1 partner seemed not to affect test-takers' speaking score as expected, although the small sample size (N=40) might have limited the generalisability of this result. The implications of this result will be discussed below in Section 6.2.2 and Section 6.2.3.

The result appears to contradict with the intelligibility benefit among test-takers from the same L1 background found in previous studies (e.g., Bent and Bradlow, 2003; Harding, 2012; Kachi, 2004; Ockey and French, 2014; Stibbard and Lee, 2006). However, this might be due to the type of task used in this study. The paired speaking tasks in this study were the collaborative tasks which required test-takers to interact in pairs. This task type seemed not to obviously reflect the test-takers' comprehension, compared to tasks such as word recognition (Stibbard and Lee, 2006), sentence recognition (Bent and Bradlow, 2003) and a word-for-word dictation task (Kachi, 2004). These tasks used in previous studies were more directly affected by the level of test-takers' comprehension and listening ability than the collaborative task as operated in this study.

6.1.1.3 Research Question 3

What are the similarities and differences in communication patterns between shared L1 pairs and non-shared L1 pairs?

Findings

The test-takers in both types of pair provided similar communication patterns related to interactive listening in terms of (1) supplying relevant vocabulary (2) demonstrating comprehension, and (3) back-channelling. Different communication patterns related to interactive listening between the shared L1 pairs and the non-shared L1 pairs concerned attempting to understand a partner completely and understanding unclear utterances.

The results from CA identified that the test-takers in both types of pair produced similar and different communication patterns related to communicative effectiveness and interactive listening behaviour during paired interaction. The listener test-takers in both types of pair provided vocabulary that their partner was searching for to complete their partner's utterance, which demonstrated their engagement with and listening comprehension regarding their partner's talk. This behaviour enabled the interaction to continue (Ducasse, 2010; Ducasse and Brown, 2009). Providing related vocabulary is considered as one type of interactive listening (Ducasse, 2010; Ducasse and Brown, 2009).

An investigation of the interactional data suggested that the test-takers in both the shared and non-shared L1 pairs demonstrated sufficient comprehension to present the evidence of their interactive listening during a conversation through relevantly responding to what their partner had said. Evidence of demonstrating comprehension is a key interactional factor for successful interaction in pairs (Ducasse, 2010; Ducasse and Brown, 2009).

Back-channelling was used by listener test-takers in both the shared and the non-shared L1 pairs as feedback to encourage a speaker to continue speaking or to let a speaker know that he/she was listening to and understanding what the speaker was saying. Two signals of back-channelling – verbal, e.g., *yeah*, *ok*, *uh huh*, and *mm*, and non-verbal, e.g., nodding the head (Ducasse, 2010; Ducasse and Brown, 2009; Shelly and Gonzalez, 2013) – were found in this study. However, producing

back-channelling did not always mean that the listener understood their partner completely (Ducasse, 2010; Ducasse and Brown, 2009). Some test-takers reported in the stimulated recall interview that they did not comprehend their partner's talk, but they just wanted their partner to keep talking (see Excerpt 18).

It is not uncommon that communication problems occur during the paired interaction between NNSs, especially with those test-takers who are from different L1 backgrounds. There was some evidence from the CA results of communication problems related to the test-takers' listening proficiency in both types of pair:

1 No response to a partner's question or speech: *back-channelling and being quiet, supplying vocabulary and being quiet (only in the non-shared L1 pairs), shifting to a new topic and being quiet;*
2 Miscomprehension;
3 Confirmation check by repeating a keyword or a question; and
4 Making a clarification request.

Based on the evidence regarding communication problems, there were three differences in communication patterns related to interactive listening between the test-takers in both types of pair: (1) causing communication problems, (2) understanding unclear utterances, and (3) misunderstanding because of having different cultural backgrounds.

The communication breakdowns seemed to relate to the limited linguistic ability of both speaker and listener test-takers and the effect of their L1 background. When communication problems occur, the test-takers usually attempt to solve them by using various explicit strategies, e.g., clarification requests, body language, background knowledge, and personal and social awareness (Hahn and Watts, 2011) in order to achieve their communication goal. However, there were some differences in solving communication problems between the test-takers in the two groups. The test-takers in the shared L1 pairs always attempted to solve communication breakdowns by using various strategies and these attempts were always successful, while the test-takers in the non-shared L1 pairs did not always try to do so and their attempts were not always successful either. In the non-shared L1 pairs, it was hard to solve all communication problems even if test-takers attempted to use various strategies. The findings of this study are incongruent with

the study of Varonis and Gass (1985) which found that NNS–NNS interaction between those from different L1 backgrounds tended to negotiate meaning more frequently than between those from the same L1 background. The results however confirmed that the pairs who did not share an L1 background had less communicative efficiency than the shared L1 pairs (van Engen et al., 2010, Varonis and Gass, 1985a).

In the shared L1 pairs, even unclear utterances could be understood easily without relying on explicit meaning negotiation, which surprised the raters because they were unintelligible to them (The same result was also found in May, 2007, 2009). This could be because shared L1 pairs may know common English mistakes that the same L1 speakers would make, and they may utilise the same L1 linguistic background knowledge (Bent and Bradlow, 2003; Hahn and Watts, 2011; Kachi, 2004) and cultural background (Kachi, 2004) to solve the communication breakdowns. The questionnaire result supported the evidence of this CA finding in terms of the test-takers being more familiar with English spoken by the shared L1 speakers than by the non-shared L1 speakers (see Section 3.4.1.1 and Appendix 11). The negative perception of English spoken by those with a different L1 in terms of pronunciation was also found in Isaacs's (2013) study, although it was expressed only by asymmetric passive test-takers.

Different cultural backgrounds could cause difficulty in understanding between the test-takers in the non-shared L1 pairs (see Excerpt 22) and could even cause communication problems. Both cultural and L1 variables seemed to be essential for the paired interactional achievement, and the findings from this study echo the importance of these two variables for spoken language used in face-to-face communication (Bachman and Palmer, 1996; Jacoby and Ochs, 1995; Young, 2000). The test-takers in the non-shared L1 pairs tended not to be enthusiastic about solving the communication problems and concentrated only on how to initiate their own talk. This might have been because they perceived that is was difficult to understand their partner's speech. It seemed that the effect of cultural and L1 backgrounds on speaking performance was more complicated than expected, so test-designers should be careful when using this format with various L1 test-takers in order to ensure that a test is tapping into the same construct between shared and

non-shared L1 pairs.

6.1.1.4 Additional interactional features of test-takers in shared and non-shared L1 pairs

Two additional interactional features were observed in the current study: the test-takers (1) provided their L1 back-channelling while listening and (2) inserted L1 words into their speech.

When the test-takers were listening to their partner, back-channelling was utilised to show their engagement with and comprehension of what their partner had been saying (see Excerpt 23). Sometimes they uttered their L1 back-channelling unconsciously. This was evidenced in Thai L1 test-takers but not in Urdu L1 test-takers. Thai back-channellings were found three times in relation to two Thai L1 test-takers. The interesting point is that L1 back-channelling occurred only when those Thai L1 test-takers were paired with a non-shared L1 partner. However, the use of L1 back-channelling seemed not to impede their interaction because their partner did not recognise it.

It was not only the Thai L1 test-takers who took the role of listener who produced their L1 back-channelling; the Thai L1 test-takers who took the role of speaker also unconsciously pronounced an L1 word (see Excerpt 24). A Thai word seemed to be unintentionally pronounced in the mental translation process.

The unconscious use of L1 back-channelling and L1 words in the non-shared L1 pairs seemed to relate to the test-takers' low level of English proficiency, their limited knowledge of using the language properly, their stress under the testing context and unfamiliarity with English spoken by different L1 speakers. Additionally, interacting with different L1 speakers seemed to require a higher cognitive demand to decode speech perception than to interact with shared L1 speakers.

6.2 Implications of the findings and contributions of the present study

The current study provides several contributions and implications for language testing research and practice. The discussion about these contributions and implications in terms of the use of the mixed-methods approach in language testing

research (Section 6.2.1), paired speaking test construct and paired speaking testing practice (Section 6.2.2), rating paired speaking tests and fairness (Section 6.2.3) and paired work and paired speaking tests in pedagogical settings (Section 6.2.4) are presented in this section.

6.2.1 The use of the mixed-methods approach in language testing research

The present study used a mixed-methods approach in order to obtain a more comprehensive understanding of the findings. While the usefulness of mixed methods, especially including CA methodology, to research paired speaking formats and group speaking formats has previously been demonstrated (e.g., Galaczi, 2014; Nakatsuhara, 2013; Van Moere, 2007) the current study has highlighted once again the usefulness of the mixed method approach by systematically and comprehensively analysing and triangulating different kinds of data.

The data in this study were listening and speaking scores, the raters' perception of the test-takers' speaking performance in stimulated recall interviews and written comments on the reasons for awarding the test-takers' speaking score, the test-takers' stimulated recall interviews and the interactional discourse data in the paired speaking formats. The most crucial contribution of this study is that it has strengthened the interdisciplinary connection between quantitative analysis, CA and other qualitative analyses by emphasising the methodological benefits of CA use to inform the relationship between test-takers' listening proficiency, their L1 and their paired speaking performance.

The present study has involved transcriptions of 40 paired sessions in total with the test-takers who were from two different L1 backgrounds (see Section 3.3.1.1 for the details of the test-takers). The sample size is relatively small, but it is still considered to be sufficient for using quantitative analysis to support the CA results. Although using quantification in CA is arguable regarding its unsuitability for interactional studies, which require a meaningful explanation and an understanding in detail of the relevant occurrence's environment (Schegloff, 1993), the researcher attempted to quantify the qualitative data as much as possible. Therefore, this study has succeeded in presenting in-depth discourse data through

the qualitative analysis of interactional features and has quantified the relevant features in order to increase generalisability and the representation of the results (Galaczi, 2014). However, as Galaczi (2014, p. 21) opines,

> In any future studies, it needs to be borne in mind that interactional data are not readily reduced to coding categories. The coding scheme used here presents a potential tool for future interactional studies, but any quantitative investigation would need to be accompanied by a thorough qualitative analysis of the discourse generated. It is in such a mixed-methods approach that the most useful insights would emerge.

Therefore, it is highly recommended that further studies are done in this area to quantify the interactional data where possible to prevent over-interpretation and increase generalisability and the representation of the results.

Since the previous literature on the impacts of test-taker characteristics on their paired speaking scores has mainly focused on the features of paired speaking discourse (e.g., Berry, 2004; Galaczi, 2014, 2004; Nakatsuhara, 2004), research on how non-native speaking test-takers with a shared and non-shared L1 interact with the target language in paired speaking tests is rarely found. Therefore, this study is systematic and comprehensive research which provides a more insightful perspective on the paired interaction discourse with the focus on the test-takers' L1 backgrounds. The analysis of the paired speaking discourse of the test-takers in the shared and non-shared L1 pairs has confirmed that a lot can be understood by using the mixed-methods approach. As shown in the results gained from the quantitative data analysis (see Sections 4.4.1 and 4.4.2), there was no difference between the test-takers' analytical scores in the shared L1 pairs and in the non-shared L1 pairs. When considering the correlation between the listening scores and the analytical scores in the paired speaking of the shared L1 pairs and the non-shared L1 pairs, the correlations were found only in the grammar and vocabulary and discourse management categories, with moderate strength. With regard to only the quantitative results, it seemed that L1 did not have an influence on the test-takers' speaking scores. However, there was some evidence in the qualitative data analysis that showed the impact of the L1 background of the test-takers and their interlocutors on their speaking performance in pairs. For example, outstanding similarities (see Section 5.1.1) and differences (see Section 5.1.2) in communication patterns related to the test-takers' interactive listening, as well as

additional interactional features (see Section 5.2) of the test-takers in the shared L1 pairs and the non-shared L1 pairs, have been found. The most interesting findings were that the test-takers in the shared L1 pairs seemed to understand their partner's message easily even if there was unclear or incorrect grammar or word use (see Section 5.1.2.2), that the shared L1 pairs tended to be more collaborative when communication breakdowns occurred and solved these problems successfully while the non-shared L1 pairs did not always attempt to repair communication breakdowns, and that Thai L1 test-takers unconsciously pronounced Thai back-channelling and words (see Sections 5.2.1 and 5.2.2) when interacting with a non-shared L1 partner. Data from the other resources – the test-takers' scores, raters' perceptions as confirmed in the stimulated recall interviews and the test-takers' stimulated recall interviews – support the CA results. Without triangulating the data from these resources, it is hard to get a comprehensive understanding of the findings. For example, in the case of the test-takers in the non-shared L1 pairs who used back-channelling to pretend to understand their partner's talk (see Excerpt 3 in Section 3.3.5.3.2), the stimulated recall interviews with the test-takers and the raters' comments helped to understand the ambiguous incidents. Therefore, it is highly recommended to utilise the mixed methods approach to study the test-takers' interaction in order to gain an in-depth understanding of the test-takers' interactive listening and interactional discourse and to generalise the results to other similar contexts.

6.2.2 The paired speaking test construct and paired speaking testing practice

This section discusses the test construct and testing practice of the paired speaking formats in light of the findings of this research.

6.2.2.1 Construct definition of paired speaking tests

The theoretical implications for research in language testing which this study has illustrated are presented in this section. As shown in Section 2.1.1, construct definitions come from three different perspectives: (1) *a trait theorist perspective* (2) *a behaviourist perspective*, and (3) *an interactionalist perspective*. The present study embraces the construct definition of the interactionalist perspective, which

views performance as a "sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts" (Chapelle, 1998, p. 43). The findings of this study have clearly shown that the characteristics of the test-takers and their partners, as well as test conditions (in this case, matching with a shared or non-shared L1 partner), do affect the test-takers' speaking performance in pairs, as discovered in previous studies (e.g., Berry, 2007; Galaczi, 2004; Jenkins, 1997, 2002; Lu, 2010; May, 2007, 2009; Nakatsuhara, 2009, 2013; Ockey, 2009, 2011; O'Sullivan, 2002, 2008). The study confirms the notion of McNamara (1997), which is that test-taker performance in pairs is related to the candidate's underlying competence and other factors. Assessing paired speaking performance is complicated and not easy; therefore, paired speaking tests should be carefully designed and administered in order to measure what the test-designer wishes to assess. These issues are also raised in the socio-cognitive framework proposed by Weir (2005). To ensure validity of the paired speaking tests, Weir's (2005) socio-cognitive framework (see Figure 2.1 in Section 2.1.2) is recommended since it includes all aspects of testing construct in practical ways.

Furthermore, this study exhibits NNS–NNS interactions in the shared L1 pairs and the non-shared L1 pairs in the context of a controlled test, which could contribute to the development of interactional competence theory (Young, 2000, p. 1), which is a theory of spoken language used in face-to-face interaction with the focus on "the structure of recurring episodes of face-to-face interaction in context, episodes that are of social and cultural significance to a community of speakers".

According to the interactionalist perspective, the definition of the construct permits test-users to generalise test-takers' performance from one context to another similar context (Chapelle, 1998). Since this study provides comprehensive interactional data related to the test-takers' L1, their listening proficiency and their speaking performance in pairs, it could be an advantageous source for the theory of interactional competence. It possibly helps with the generalisability of the interactional performance of NNS–NNS to a similar context by providing a more accurate picture.

6.2.2.2 Paired speaking testing practice

According to McNamara (1996), assessing test-takers' speaking performance in the paired formats is related to the test-taker's underlying competence and other co-constructing factors (see Figure 1.2). This study provides the empirical data to inform a test-designer's decision about whether test-takers' L1 and their partners' L1 backgrounds should be included in the paired speaking test construct.

Two types of threat to test validity that test-designers should bear in mind were described in Chapter 2 (Section 2.1.2), i.e., *construct under-representation and construct-irrelevant variance* (Messick, 1989). One of the underlying aims of this study was to offer some evidence to help understand how to conceptualise the L1 backgrounds of test-takers and their partners to reduce the risk of the two types of threat to the test validity. This study has illustrated the effect of test-takers' L1 and their listening proficiency on the paired interactions, and it highlighted the importance of interactive listening skills in relation to these variables for effective communication in the paired speaking tests. Based on the results of this study, it can be suggested that test-takers' L1 variable should be a part of the paired test construct if the test is to be a good predictor of test-takers' achievement in interactive conversation. While the L1 variable may affect test-takers' communication patterns and advantage and disadvantage different test-takers, it still contributes to create an interactional environment in which test-takers can display their interactive listening skills and interactional competence. It is also important to keep in mind that the L1 background variable is also the variable that affects conversation between shared L1 partners and non-shared L1 partners in the real world.

Nevertheless, for the paired speaking test in the setting of test-takers from multi-L1 backgrounds, test-designers should consider the implication of forming shared L1 pairs and non-shared L1 pairs, in terms of test fairness. Test-takers may gain some advantages or disadvantages in their paired speaking interaction due to their partners' L1 backgrounds. As shown in this study, test-takers in the shared L1 pairs seemed to gain some advantages in terms of an increased level of intelligibility, interaction involving less stress, more collaborative interaction and more successful repairs of communication problems than the test-takers in the non-

shared L1 pairs. On the other hand, test-takers in the shared L1 pairs might be disadvantaged in terms of careless use of grammar and vocabulary and pronunciation because they seemed to understand each other easily (May, 2007, 2009) even when pronouncing an unclear utterance or using an incorrect word or incorrect grammar (see Section 5.1.2.2), which will be awarded negative scores by raters. In this sense, test-designers should consider which types of pairing would be the best to measure the test construct in their contexts.

Pairing with a shared and non-shared L1 partner involves advantages and disadvantages in different aspects, though the paring methods might not affect resulting scores; therefore, if time and resources allow, utilising both types of pair for all test-takers in the testing context of paired speaking formats is recommended. However, it is likely to be unrealistic to administer both types of pair for all test-takers in practice, especially in high-stakes testing contexts. It is not cost-effective or time-efficient in large-scale testing contexts, and it will also be very difficult to find a partner from the same L1 background for all test-takers, especially when their L1 is not a common one. Therefore, for any testing contexts in which students with various L1 backgrounds are involved, it is recommended that Swain's (1983) concept of "*bias for best*" be adopted in paired speaking test practice in order to create testing conditions to allow for the best performance of test-takers (Fox, 2004). The examiners should allow test-takers to select the partner with whom they think they can perform their best in paired interaction. This was also suggested by O'Sullivan (2002) as a way to address possible effects of test-taker characteristics in paired tests.

6.2.3 Rating paired speaking tests and fairness

This section presents the implications for rating paired speaking interaction, which are perceived in the present study as follows: (1) further development of the scale for the interactive communication category for the paired speaking tests and (2) fairness in awarding a score for each test-taker's performance in pairs.

6.2.3.1 Further development of the scale for the interactive communication category for the paired speaking tests

The CA results have shown that the paired speaking format of FCE has great potential to elicit test-takers' interactional behaviour and are appropriate for assessing the test-takers' interactive communication proficiency. However, as presented in Table 6.1 below, the FCE interactive communication scale does not mention interactive listening skills explicitly, although we can assume from the phrase "*responds appropriately*" that test-takers have to have a sufficient level of proficiency in listening, interactive listening and speaking skills in order to respond to their partner's message appropriately.

Table 6.1: The FCE speaking criteria for the interactive communication category (UCLES, 2015, p. 82)

Band	Interactive communication
5	Initiates and responds appropriately, linking contributions to those of other speakers. Maintains and develops the interaction and negotiates towards an outcome.
4	*Performance shares features of Bands 3 and 5.*
3	Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.
2	*Performance shares features of Bands 1 and 3.*
1	Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.
0	*Performance below Band 1.*

"*Linking contributions to those of other speakers*" also requires a degree of listening proficiency to decode their partner's speech and to produce their own spoken contributions (Field, 2011; Galaczi, 2014). Whenever a communication problem occurs, they are required to negotiate meaning to solve that breakdown, which helps the interaction continue (e.g., Excerpts 15 and 16). As shown in the results of this study, listening proficiency was important for the paired interaction (moderate statistical correlations between test-takers' listening scores and paired speaking scores in grammar and vocabulary and those in discourse management), for the non-shared L1 pairs (moderate correlations between test-takers' listening scores and paired speaking scores in grammar and vocabulary and discourse management in non-shared L1 pairs). However, there was no statistical correlation between listening scores and interactive communication scores in the present study. The

results gained from CA show that interactive listening ability, i.e., *supplying relevant vocabulary*, *demonstrating comprehension* and *back-channelling*, is important for the successful interaction of both types of pairing, as illustrated in Section 5.1.1. The importance of interactive listening skills for successful paired interaction is also highlighted in the literature (e.g., Ducasse, 2010; Ducasse and Brown, 2009; Galaczi, 2014). Therefore, it is recommended that the criterion of interactive communication should refer to interactive listening ability as one of the factors needed for interactional achievement, and this is also suggested by Galaczi (2014) and Ducasse (2010). The scale should clearly specify what types of response can be considered as "*appropriate*", i.e., whether short responses, back-channelling use and non-verbal responses are acceptable as appropriate responses that help raters have clear guideline on how to score the interaction ability. Indeed, this study found that responding by using these features did not always mean that the test-takers really understood their partner, and they were used to encourage their partner to continue speaking until they understood their partner and could switch from being in a listener to in a speaker role (Ducasse, 2010; Ducasse and Brown, 2009), as illustrated in Excerpt 18. Moreover, there is a case that the listener test-takers did not understand their partner's speech completely but pretended to listen to and understand their partner. An explicit example of this is shown in Excerpt 3. This evidence was also found in the studies of Ducasse (2010) Ducasse and Brown (2009). Hence, raters should not only rely on the use of back-channelling to assess test-takers' comprehension. They should evaluate how much they are able to extend or elaborate their partner's speech when they change from a listener to speaker role, as an indicator of interactive listening skills.

Based on the results of the present study and the reviewed literature, the following descriptions for interactive communication related to interactive listening ability can be suggested as guidelines for the criteria:

1 Test-takers are awarded positive scores in the following performances:
    a. They are able to use back-channelling and non-verbal signals effectively and appropriately in presenting that they are listening to, understanding and supporting their partner to talk;

b. They are able to manage communication breakdowns by using various strategies, e.g., negotiating meaning and/or making clarification requests and confirmation checks; and
c. They are able to help their partner when he/she has difficulty in searching for appropriate vocabulary or presenting his/her ideas speaking.

2 Test-takers are awarded negative scores when:

a. They do not attempt to negotiate comprehension in order to solve communication breakdowns by requesting clarification;
b. They do not assist their partner to continue a conversation when he/she faces difficulty in delivery of an idea;
c. They do not help their partner to find an appropriate word when it is obvious that he/she is searching for it; and
d. They do not respond or respond with a short answer or minimal acknowledgement but are not able to extend or develop their partner's ideas.

These suggestions confirm what other related studies have also proposed, such as Ducasse (2010) and Ducasse and Brown (2009), Galaczi (2014) and May (2009).

Furthermore, rater training also needs to address how to treat speech that was intelligible to a paired partner but not to a rater, L1 back-channelling and L1-influenced words are intelligible to the test-takers in the shared L1 pairs, and the goal of the interactive communication is that the test-takers are able to understand each other's talk, initiate and respond to each other appropriately, and maintain and develop the interaction. It is, in this case, the interaction between the test-takers not the raters that matters. How to deal with what is unintelligible to raters but intelligible to the paired test-takers, as also questioned in May (2007), needs to be discussed in light of the concept of World Englishes and the test construct to be measured in each specific test.

6.2.3.2 Fairness in rating an individual's performance in pairs

Even though the paired speaking tests can elicit various interactional features from the test-takers and their construct is in accordance with the interactionalist

perspective, the fairness issue still remains unsolved when it comes to rating test-takers' speaking performance in pairs. As stated earlier, the characteristics of the test-takers' partner unavoidably affect the test performance. Even if these variables can be part of the test construct, how to evaluate test-takers' performance fairly while their performance is co-constructed with their partner's performance is worth consideration. The finding of this study was encouraging in this respect. The score comparison between shared and non-shared L1 pairs suggested that the paring did not significantly affect score outcomes. That is, whether test-takers are paired with shared or non-shared partners is unlikely to affect their scores.

Due to the difficulty in rating test-takers' interactional effectiveness, May (2007) suggests sharing the score for interactional effectiveness of test-takers' performance in pairs and awarding scores in other categories separately. However, as mentioned in Chapter 3, Nakatsuhara (2009) argues that sharing scores does not always guarantee fairness in an assessment, since there are some cases in which test-takers fail to get quiet members to cooperate in the interaction even though they try very hard to scaffold others' participation. This view is in line with the current FCE practice, and this study employed the individual scoring method. The findings of this study also supported this basis. For example, there was some evidence of asymmetrical interaction (e.g., Excerpt 14), which obviously presented test-takers' trying to encourage their partner to talk, but they did not succeed. Hence, it seems unfair for them to be penalised for interactive communication. For this reason, it is suggested that test-takers' performance in pairs should be rated separately. To overcome the difficulty of separate scoring, a clear and more detailed definition of the interactive communication scale, as suggested earlier, would be beneficial.

The other possibility in relation to awarding an interactional communication score and other scores fairly is to assess test-takers in various speaking test formats to reflect their real speaking proficiency as much as possible. For example, the Cambridge ESOL uses this approach in its suite of English language examinations in which all test-takers have to do a monologic speaking test format, a two-way interaction between test-takers (paired speaking format) and a three-way interaction with the other test-taker and an examiner.

6.2.4 Paired work and paired speaking tests in pedagogical settings

As stated in Section 2.2, the paired formats could elicit rich language functions from test-takers, and the test-takers provided positive feedback on the paired speaking formats. Therefore, it is recommended that the paired speaking formats be utilised in classroom settings in order to develop the linguistic and interactional competence of language learners. In classroom contexts where there are only learners from the same L1 background, they rarely use L2 to communicate with each other in their real life outside the classroom. As reviewed in Chapter 2 (Section 2.2.1), it has been reported that the paired speaking formats can foster positive washback to L2 classroom (e.g., Együd and Glover, 2001; Jones, 2007; May, 2000). The formats can help them speak using various language functions and encourage them to use L2 to communicate with each other with confidence. This can help them develop their skill in interactive communication.

In multilingual classroom assessment contexts, it is recommended that a teacher should conduct a paired speaking test by pairing students with different L1 partners at different times (Luoma, 2004). In the real world, NNS and NNS from the same L1 background rarely communicate with each other in L2; therefore, pairing students from different L1 backgrounds will be an advantage for the students in terms of being more confident and familiar with other accents when they have to communicate with other NNS outside class. Besides, the findings of this study suggested that students would need to practice how to negotiate meaning to solve communication breakdowns in non-shared L1 pairs. Teachers should teach students how to negotiate meaning when facing communication breakdowns to reach an interaction goal, and encourage them to initiate meaning negotiation during non-shared L1 interaction. Additionally, if possible, teachers could also provide them with opportunities to experience paired interaction with a shared L1 partner, and ask them to compare their interactions in shared and non-shared L1 pairs. In doing this, teachers will assist students to learn how to communicate with both shared and non-shared L1 partners effectively, and how to improve their interactional skills.

It is also important for teachers to raise students' awareness about interactional patterns, especially by explaining unsuccessful interactional patterns:

a*symmetric* and *parallel interactions*. By showing videos or transcripts of paired tests (as presented in this study), teachers can demonstrate how a dominant test-taker is judged as lacking interactional skills, while a passive test-taker is viewed as having limited speaking ability. They can also point out that non-shared L1 pairs are more likely to end up with parallel interaction patterns, if learners do not make constant efforts to listen to the partner and develop and expand on the partner's idea. Examples such as Excerpts 15 and 18 would be useful, which was called non-interaction by the raters because the test-takers initiated their own contributions without extending their partner's ideas. Raising students' awareness for successful and unsuccessful interactional patterns would help students interact with people from various L1 backgrounds effectively and with more confidence.

6.3 Limitations of the study and directions for further study

Although the current study provides various contributions and implications, there are some limitations in terms of selecting participants and generalisability.

Firstly, due to practical constraints, it was difficult to find the two L1 groups which consisted of participants from the same circle of Kachru's (1998) model of the three concentric circles of English. However, the researcher attempted to control the possible effect of this difficulty on participants' English proficiency levels by selecting participants with similar speaking and listening proficiency levels based on their IELTS band scores. Short interviews with the participants from the outer circle of English (Urdu L1 speakers) about their use of English in their daily life were also used at the participant selection stage. Even though the researcher did her best to control the variable which might affect the test-takers' interactional performance, there might be some implicit effects which might have confounded the results of this study. Hence, it is recommended that this study is replicated with test-takers with other L1s who are in the same concentric circle of English and have the same speaking and listening proficiency level.

Secondly, the results of this research with test-takers from Urdu and Thai L1 backgrounds might be not generalisable to other L1 contexts because people from different cultural and L1 backgrounds have different interactional patterns (Lu, 2010). For the benefit of generalisability, further studies should investigate the

effect of test-takers' L1 and their listening proficiency on their speaking performance in pairs or in groups with test-takers from other cultural and L1 backgrounds. One of the communication breakdowns which occurred in one of the non-shared L1 pairs in this study resulted from the test-takers' different cultural backgrounds (see Excerpt 22). Hence, studies on the effect of different cultural backgrounds of test-takers on their interactional communication are required. In such studies, great care should be taken with the topics used to elicit the interactional competence of test-takers from different cultural and L1 backgrounds in order to prevent some test-takers from gaining advantages or disadvantages from the topic used.

Additionally, this study showed some findings related to communication problems which occurred during interaction between the test-takers in the shared L1 pairs and non-shared L1 pairs. Communication problems occurred more frequently in the non-shared L1 pairs than in the shared L1 pairs. The test-takers in the shared L1 pairs always helped each other to solve a communication breakdown when it occurred, while the test-takers in the non-shared L1 pairs did not always attempt to do so. Surprisingly, the test-takers in the shared L1 pairs were able to solve all communication problems by utilising various strategies, e.g., clarifying a question, encouraging a partner to speak, gesturing, checking confirmation, while the test-takers in the non-shared L1 pairs merely attempted to use those strategies (see Section 5.1.2.1 and Table 5.7). Therefore, further studies should focus on communication problems which occur in the NNS–NNS interaction in the shared L1 pairs and non-shared L1 pairs, and investigate further how they solve those communication breakdowns in more details. The results of such studies will be highly beneficial in gaining a better understanding of NNS–NNS interaction and in facilitating more effective communication by both shared and non-shared L1 speakers who use English as a communication tool.

In addition, this study has used only the FCE collaborative task to assess the test-takers' speaking performance in pairs. For further studies, the use of different task types, such as an information exchange task, is recommended in order to see whether the same results can be obtained with different types of task. Studies with a bigger sample size are also recommended to obtain more generalisable results.

6.4 Final thoughts

As stated early in Chapter 1, English is used in international contexts, and users and learners of the English language who are in different regions speak different varieties of English. Understanding how well English is used as a tool of communication by NNSs who have the same and different L1 backgrounds and how they succeed in achieving their interaction goals more effectively has important implications in the international context. A mixed-methods approach has been utilised in this study to shed light on the interactive communication between NNS and NNS in paired speaking tasks. Although the findings of this study have highlighted the complex nature of paired interaction co-constructed by paired test-takers, they have also contributed to a better understanding of the interaction between the test-takers in the shared L1 pairs and the non-shared L1 pairs. As such, it is hoped that this thesis has offered some research guidelines for future researchers who are interested in studying paired NNS interaction in terms of their L1 backgrounds. Additionally, it is also hoped that the findings of this study will help international examination boards which use a paired format and teachers working in the classroom teaching English to international students to make an informed decision about their pairing methods in order to enhance fairness to test-takers/learners and improve the effectiveness of their teaching.

References

Algethami, G., Ingram, J. and Nguyen, T. (2011). The interlanguage speech intelligibility benefit: The case of Arabic-accented English. In. J. Levis and K. LeVelle, editors, *Proceedings of the 2nd Pronunciation in Second Language Learning and Teaching Conference*, Sept. 2010. (pp. 30-42), Ames, IA: Iowa State University.

Atkinson, J.M. and Heritage, J. (1984). *Structures of social action.* Cambridge: Cambridge University Press.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. and Palmer, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bennett, R. (2012). Is linguistic ability variation in paired oral language testing problematic? *ELT Journal* 66(3), pp. 337-346.

Bent, T. and Bradlow, A.R. (2003). The interlanguage speech intelligibility benefit, *Journal of the Acoustical Society of America*, 114(3), pp. 1600-1610.

Berry, V. (1993). Personality characteristics as a potential source of language test bias. In A. Huhta, K. Sajavaara and S. Takala, editors, *Language testing: new openings* (pp. 115-124). Jyvaskyla, Finland: Institute for Educational Research.

Berry, V. (1997). Ethical considerations when assessing oral proficiency in pairs. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma, editors, *Current developments in language testing* (pp. 107-123). Jyvaskyla: Jyvaskyla University Press.

Berry, V. (2004). *A study of the interaction between individual learner differences and oral performance test facets*. Unpublished PhD. Thesis. University of London.

Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter Lang.

Brooks, L. (2009). *Paired oral proficiency testing: Challenges but mostly opportunities*. Paper presened at AILA, Germany, 26/08/2008.

Brown, A. (2003). Interview variation and the co-construction of spekaing proficiency. *Language Testing* 20(1), pp. 1-25.

Brown, H.D. (2004). *Language Assessment: Principle and and classroom practice*. New York: Longman.

Brown, J.D. (2005). *Testing in language programs: a comprehensive guide to English language assessment.* New York: McGraw-Hill.

Brown, G. and Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Buck, G. (2001). *Assessing listening*. New York: Cambridge University Press.

Cambridge ESOL, (2008). *Cambridge Preliminary English Test*. Cambridge: Cambridge University Press.

Cambridge ESOL, (2009). *Cambridge First Certificate in English*. Cambridge: Cambridge University Press.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards and R. W. Schmidt, editors, *Language and communication* (pp. 2-27). London: Longman.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1), pp. 1-47.

Carey, M.D., Mannell, R.H. and Dunn, P.K. (2011). Does rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, 28(2), pp. 201-219. doi: 10.1177/0265532210393704

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing* 20(4), pp. 369-383.

Chapelle, C.A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. Cohen, editors, *Inferences between second language acquisition and language testing research* (pp. 32-70). New York: Cambridge University Press.

Cohen, J. (1988). *Statistical power for analysis for the behavioural sciences (2nd edition).* New York: Academic Press.

Coolican, H. (2009). *Research Methods and Statistics in Psychology (5th edition).* New York: Routledge.

Council of Europe (2001). *Common European Framework of Reference for*

Languages: learning, teaching, assessment. Cambridge: Cambridge University Press.

Creswell, J.W. and Plano Clark, V.L. (2011). *Designing and conducting mixed methods research (2nd edition)*. California: Sage Publications.

Crystal, D. (1997). *English as a global language (2nd edition)*. Cambridge: Cambridge University Press.

Csépes, I. (2009). *Measuring oral proficiency through paired-task performance*. Frankfurt am Main: Peter Lang.

Davies, A., et al. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing* 26(3), pp. 367-396.

Derwing, T.M. and Munro, M.J. (1997). Accent, intelligibility and comprehensibility: evidence from four L1s. *Studies in Second Language Acquisition* 19(1), pp. 1-16.

Derwing, T.M. and Munro, M.J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching* 42, pp. 476-490.

Dörnyei, Z. (2007). Creating a motivating classroom environment. In J. Cummins and C. Davison (Eds.), *International handbook of English language teaching* (Vol. 2, pp. 719-731). New York: Springer.

Douglas, D. (1998). Testing methods in context-based second language research. In L.F. Bachman and A. Cohen, editors, *Interfaces between second lanaguage acquisition and lanague testing research* (pp. 141-155). Cambridge: Cambridge University Press.

Ducasse, A.M. (2010). *Interaction in paired oral proficiency assessment in Spanish: Rater and candidate input into evidence based scale development and construct definition*. Frankfurt: Peter Lang.

Ducasse, A.M. and Brown, A. (2009). Assessing paired orals: raters' orientation to Interaction. *Language Testing* 26(3), pp. 423-443.

Educational Testing Service (1989). *Oral proficiency testing manual.* Princeton, NJ: Educational Testing Service.

Együd, G. and Glover, P. (2001). Oral testing in pairs: a secondary school

perspective. *English Language Teaching Journal* 55(1), pp. 70-76.

Fayer, J.M. and Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning* 37, pp. 313–326.

ffrench, A. (2003). The change process at the paper level. Paper 5, Speaking. In C. J. Weir and M. Milanovic, editors, *Continuity and innovation: revising the Cambridge Proficiency in English examination 1913-2002* (pp. 367-471). Cambridge: Cambridge University Press.

Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top-down? *TESOL Quarterly*, *32*(3), 363–377.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*(3), 399–423.

Field, J. (2008). *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Field, J. (2011). Cognitive validity. In L. Taylor, editor, *Examining Speaking* (pp. 65-111). Cambridge: Cambridge University Press.

Foot, M.C. (1999). Relaxing in pairs. *ELT Journal* 53(1), pp. 36-41.

Fox, J. (2004). Biasing for the Best in Language Testing and Learning: An Interview with Merrill Swain. *Language Assessment Quarterly* 1(4), pp. 235-251.

Fulcher, G. (2003). *Testing second language speaking*. London: Longman.

Galaczi, E. (2004). *Peer-peer interaction in a paired speaking test: The case of the First Certificate in English.* Unpublished PhD thesis, Columbia University.

Galaczi, E.D. (2014). Interactional competence across proficiency levels: How do learners managing interaction in paired speaking tests? *Applied Linguistics* 35(5), pp. 553-574.

Gass, S.M. and Varonis, E.M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning* 34, pp. 65–89.

Gass, S.M. and Mackey, A (2000). *Stimulated recall methodology in second language research*. New Jersey: Lawrence Erlbaum.

Gass, S.M. and Varonis, E.M. (1991). Miscommunication in non-native speaker discourse. In N. Coupland, H. Giles and J. M. Wiemann, editors, *"Miscommunication" and problematic talk* (pp. 121-145). California: Sage.

Geranpayeh, A. (2007). Using Structural Equation Modelling to facilitate the revision of high stakes testing: the case of CAE. *Research Notes* 30, pp. 8-12.

Graham, M., Milanowski, A. and Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. on-line, accessed on 10/01/2014 at http://www.files.eric.ed.gov/fulltext/ED532068pdf

Hahn, L. and Watts, P. (2011). (Un)Intelligibility tales. In Levis, J. and LeVelle, K.. editors, *Proceeding of the 2nd pronunciation in second language learning and teaching conference* (pp. 17-29). Ames, I.A.: Iowa State University.

Hamp-Lyons, L. and Davies, A. (2008). The Englishes of English tests: bias revisited. *World Englishes* 27(1), pp. 26-39.

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: DIF perspective. *Language Testing* 29(2), pp. 163-180.

Hinton, P.R., Brownlow, C., McMurray, I. and Cozens, B. (2004). *SPSS Explained*. New York: Routledge.

Hutchby, I. and Wooffitt, R. (1998). *Conversation Analysis*. Malden: Blackwell Publishers Inc.

Isaacs, T. (2010). *Issues and arguments in the measurement of second language pronunciation*. Unpublished PhD thesis, McGill University, Montreal.

Isaacs, T. (2013). International engineering graduate students' interactional patterns on a paired speaking test: Interlocutors' perspectives. In K. McDonough and A. Mackey, editors, *Second language interaction in diverse educational settings* (pp. 227-245). Amsterdam: John Benjamins.

Isaacs, T. (2014). Assessing pronunciation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 140–155). Hoboken, NJ: Wiley-Blackwell.

Isaacs, T., Trofimovich, P., Yu, G., and Chereau, B.M (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS research reports online series, 4.*

Itakura, H. (2001). Describing conversational dominance. *Journal of Pragmatics* 33, pp. 1859-1880.

Iwashita, N. (1998). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing* 5(2), pp. 51-65.

Jacoby, S. and Ochs, E. (1995). Co-construction: an introduction. *Research on Language and Social Interaction* 28(3), pp. 117-183.

Jaiyote, S. (2015). *The relationship between test-takers' L1, their listening proficiency and performance on paired speaking tests*. on-line, accessed on 10/01/2016 at https://www.britishcouncil.org/sites/default/files/the_relationship_between_test-takers_l1.pdf

Jenkins, J. (1997). Testing pronunciation in communicative exams. In M. Vaughan-Rees, editor, *A special issue of Speak out! Bringing together the interests of the IATEFL pronunciation and testing sigs* (pp 7-11).

Jenkins, J. (2000). *The Phonology of English as an International Language*. Oxford: Oxford University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an International Language. *Applied Linguistics* 23, pp. 83-103.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *ELT Journal* 60 (1), pp. 42-50.

Jenkins, J. (2007). *English as a Lingua Franca: Attitude and Identity*. Oxford: Oxford University Press.

Johnson, M. (1997). *What kind of speech event is the Oral Proficiency Interview: Problems of construct validity*. Unpublished Georgetown University Dissertation.

Johnson, M. (2000). Interaction in the oral proficiency interview: problems of validity. *Pragmatics* 10(2), pp. 215-231.

Johnson, M. (2010). *The art of non-conversation: A re-examination of the validity of the oral proficiency Interview*. Maryland: Yale University Press.

Johnson, M. and Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? In R. Young and A. W. He, editors, *Talking and testing: discourse approaches to the assessment of oral proficiency* (pp. 27-51). Amsterdam, Philadelphia: John Benjamins.

Jones, E.E. and Gerard, H.B. (1967). *Foundations of social psychology*. New York: John Wiley & Sons

Jones, L. (2007). *The student-centred classroom*. New York: Cambridge University Press.

Kachi, R. (2004). *Factors predicting native and non-native listeners' evaluative reactions to Japanese English*. PhD thesis, Ohio State University.

Kachru, B.B. (1992). World Englishes: approaches, issues and resources. Language Teaching 25(1), pp.1-14.

Kachru, B.B. (1998). English as an Asian Language. *Links & Letters* 5, pp. 89-108.

Kachru, B.B. (2005). *Asian Englishes Today: Asian Englishes beyond the Canon*. Hong Kong: Hong Kong University Press.

Kachru, Y. and Smith, L.E. (2008). *Cultures, Contexts and World Englishes*. New York: Routledge.

Kenkel, J.M. and Tucker, R.W. (1989). Evaluation of institutionalized varieties of English and its implications for placement and pedagogy. *World Englishes* 8(2), pp. 201-214.

Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal* 70(4), pp. 366-372.

Lazaraton, A. (1992). The structural organization of a language interview: a conversation analytic perspective. *System* 20, pp. 373-386.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 12(2), pp. 151-172.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Levinson, S.C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Liao, S.O., Hunt, E.A. and Chen, W. (2010). Comparison between inter-rater reliability and inter-rater agreement in performance assessment. *Ann Acad Med Singapore* 39(8), pp. 613-618.

Lu, Y. (2010). Chinese EFL learners' discoursal performance in the Cambridge ESOL FCE Speaking Test. In L. Cheng and A. Curtis, editors, *English Language Assessment and the Chinese Learner* (pp. 175-189), New York: Routledge.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Major, R.C., Fitzmaurice, S.F., Bunta, F. and Balasubramanian, C. (2002) The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly* 36(2), pp. 173-190.

May, L. (2000). Assessment of oral proficiency in EAP programs: A case for pair interaction. *Language and Communication Review* 9(1), pp. 13–19.

May, L. (2007). *Interaction in a paired speaking test: the rater's perspective*. Unpublished PhD thesis, the University of Melbourne.

May L. (2009). Co-constructed interaction in a paired speaking test: the rater's perspective. *Language Testing* 26(3), pp. 397-421.

McCarthy, M. (1991). *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press.

McNamara, T.F. (1996). *Measuring second language performance*. Harlow: Longman.

McNamara, T.F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics* 18(4), pp. 446-466.

McNamara, T.F. (2000). *Language Testing*. New York: Oxford University Press.

McNamara, T.F. (2003). Knowing your limits: Conceptual and policy constraints on evidence centred test design in language testing research. *Measurement: Interdisciplinary Research and Perspectives* 1(1), pp. 85–87.

McNamara, T.F., Hill, K. and May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics* 22, pp. 221-242.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin* 89(3), pp. 575-588.

Messick, S. (1989). Validity. In R.L. Linn, editor, *Educational measurement (3rd edition)* London, NY: McMillan.

Munro, M.J. and Derwing, T.M. (2015). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation* 1(1), pp. 11-42.

Nakatsuhara, F. (2004). *An investigation into conversational style in paired speaking tests*. Unpublished MA thesis, the University of Essex.

Nakatsuhara, F. (2006). The impact of proficiency-level on conversational styles in

paired speaking tests. *Cambridge ESOL Research Notes* 25, pp. 15-20.

Nakatsuhara, F. (2009). *Conversational styles in group oral tests: How is the conversation co-constructed?* Unpublished PhD thesis, the University of Essex.

Nakatsuhara, F. (2011). The relationship between test-takers' listening proficiency and their performance on the IELTS Speaking Test. *IELTS Research Reports* 12, pp. 1-50.

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Frankfurt am Main: Peter Lang.

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal* 59(4), pp. 287-296.

Ockey, G.J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing* 26(2), pp. 161-186.

Ockey, G.J. (2011). Self-consciousness and assertiveness as explanatory variables of L2 oral ability: a latent variable approach. *Language Learning* 61(3), pp. 968-989.

Ockey, G.J. and French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. *Language Testing* 3, pp. 1-24.

O'Sullivan, B. (1995). *Oral language testing: does the age of the interlocutor make a difference?* Unpublished MA dissertation, University of Reading.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System* 28 (3), pp. 373-386.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing* 19(3), pp. 277-295.

O'Sullivan, B. (2008). *Modelling Performance in Tests of Spoken Language*. Frankfurt: Peter Lang.

O'Sullivan, B. (2013). Assessing Speaking. *The Companion to Language Assessment* 2(9), pp. 156-171.

O'Sullivan, B. and Lu, Y. (2002). The impact on candidate language of examiner deviation from a set interlocutor frame in the IELTS Speaking Test. *IELTS Research Reports* 6, pp. 91-117.

O'Sullivan, B. and Weir, C. (2011). Test development and validation. In:

O'Sullivan, B. (ed.) *Language testing: theories and practices*. Basingstoke: Palgrave 'Advances in Language and Linguistics' series Hampshire, pp.13-32.

O'Sullivan, B., Weir, C. and Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing* 19(1), pp. 33-56.

O'Sullivan, B. and Saville, N. (2000). *Using observation checklists to validate speaking-test tasks*. Paper presented at the Language Testing Research Colloquium, Vancouver, March 2000.

Phootirat, P. (2012). *Register variation in Thai-English interphonology: The contrast of /r/ and /l/*. Unpublished PhD thesis, the University of Wisconsin-Milwaukee.

Plough, I.C., MacMillan, F., and O'Connell, S.P. (2011). Changing Tasks … Changing Evidence: A Comparative Study of Two Speaking Proficiency Tests. In Granena, G., Koeth, J., Lee-Ellis, S., Lukyanchenko, A., Botana, G.P. and Rhoades, E., editors, *Selected Proceedings of the 2010 Second Language Research Forum* pp. 91-104. Somerville, MA: Cascadilla Proceedings Project.

Porter, P.A. (1986). How learners talk to each other: input and interaction in task-centred discussion. In R.R. Day, editor, *Talking to learn: conversation in second language acquisition* (pp. 200-224). Cambridge, Massachusetts: Newburry House Publishers.

Rahman, T. (1990). *Pakistani English*. on-line accessed on 09/01/2015 at http://www.tariqrahman.net/content/pakenglish.pdf.

Sacks, H., Schegloff, E.A. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4), pp. 696-735.

Schegloff, E.A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction* 26(1), pp. 99-128.

Shackle, C. (2001). Speakers of South Asian language. In M. Swan and B. Smith, editors, *Learner English: A teacher's guide to interference and other problems (2nd edition)* (pp. 227-243).Cambridge: Cambridge University Press.

Shelly, L. and Gonzalez, F. (2013). Back Channelling: Function of Back

Channelling and L1 Effects on Back Channelling in L2. *Linguistic Portfolios* 2, pp. 98-108.

Shiotsu, T. and Weir, C.J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing* 24(1), pp. 99-128.

Smith, M.S. and Kellerman, E. (1986). Crosslinguistic influence in second language acquisition: An introduction. In: Kellerman, E. and Smith, M.S. (eds.) *Crosslinguistic influence in second language acquisition*. Oxford: Pergamon Press, pp. 1-9.

Smyth, D. (2001). Thai speakers. In M. Swan and B. Smith, editors, *Learner English: A teacher's guide to interference and other problems (2nd edition)* (pp. 343-356). Cambridge: Cambridge University Press.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning* 52(1), pp. 119-158.

Stibbard, R.M. and Lee, J.I. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *Acoustical Society of America* 120, pp. 433-442.

Swain, M. (1983). Large-scale communicative language testing: A case studying. *Language Learning and Communication* 2, pp. 133-142.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*18(3), pp. 275-302.

Taylor, L. (2001). The paired speaking test format: recent studies. *Cambridge ESOL Research Notes 6*, pp. 15-17.

Taylor, L., editor (2011). *Examining speaking research and practice in assessing second language speaking. Studies in Language Testing 30*. Cambridge: UCLES/Cambridge University Press.

Taylor, L. and Wiggleworth, G. (2009). Are two heads better than one? Paper work in L2 assessment contexts. *Language Testing* 26(3), pp. 325-339.

Ten Have, P. (1999). *Doing conversation analysis: A practical guide*. Surrey: Sage.

Thompson, I. (2013). About World English. on-line accessed on 01/01/2015 at: http://aboutworldlanguages.com.

UCLES (2009). *First Certificate in English: Handbook for teachers*, on-line accessed on 16/09/12 at: https://www.lttc.ntu.edu.tw/cambridge/ms/handbook/fce/fce_handbook.pdf

UCLES (2012). *First Certificate in English: Handbook for teachers*, on-line accessed on 10/06/2013 at: http://www.cambridgeenglish.org/images/caehandbook-for-teachers.pdf.

UCLES (2014). Cambridge Preliminary English Test (PET): Handbook for teachers, on-line accessed on 5/10/2015 at: http://www.cambridgeenglish.org/images/168150-cambridge-english-preliminary-teachers-handbook.pdf

UCLES (2015). *First Certificate in English: Handbook for teachers*, on-line accessed on 02/01/2015 at: http://www.cambridgeenglish.org/images/cambridge-english-first-handbook-2015.pdf.

Van Engen, K.J., Baese-Berk, M., Baker, R.E., Choi, A., Kim, M. and Bradlow, A. R. (2010). The Wildcat Corpus of native- and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech* 53(4), pp. 510-540.

Van Lier, L. (1989). Reeling, writing, drawing, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly* 23(3), pp. 489-508.

Van Moere, A. (2007). Group oral test: How does task affect candidate performance and test score? Unpublished PhD thesis, the University of Lancaster.

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics.* 24, pp. 3-25.

Varonis, E.M. and Gass, S. (1985a). Non-native/Non-native Conversations: A Model for Negotiation of Meaning. *Applied Linguistics*. 6(1), pp. 71-90.

Varonis, E.M. and Gass, S. (1985b). Miscommunication in Native/Non-native Conversation. *Language in Society*. 14 (3), pp. 327-343.

Weir, C. (1993). *Understanding and Developing Language Tests*. Hemel Hemstead: Prentice Hall.

Weir, C.J. (2005). *Language Testing and Validation*. New York: Palgrave Macmillan.

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency*. Ann Arbor, MI: University of

Michigan Press.

Winke, P., Gass, S. and Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30, pp. 231-252.

Xiaoxian, G. and Yan, J. (2010). Interactive listening: Construct definition and operationalization in tests of English as a Foreign Language. *Chinese Journal of Applied Linguistics* 33(6), pp. 16-39.

Young, R. (2000). *Interactional competence: challenges for validity*. paper presented at a joint symposium on 'Interdisciplinary interface with language testing', held at the Annual Meeting of the American Association for Applied Linguistics and the Language Testing Research Colloquium, 11/11/2014, Vancouver, British Columbia, Canada, available on-line http://www.wisc.edu/english/rfyoung/IC_C4V.Paper.PDF.

Young, R.F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel, editor, *Handbook of research in second language teaching and learning* (pp. 426-443). New York, NY: Routledge.

Young, R. (2014). What is interactional competence? *AL Forum: The Newsletter of the Applied Linguistics Section of TESOL*.

Young, R. and Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition* 14, pp. 403-424.

Appendices

Appendix 1

Questionnaire

Please fill in <u>all sections</u> of this questionnaire.

1) About Yourself

Name: .....

Student ID number: .....

Email:.....

Gender: Male / Female (please circle) Age:.....years old

Country of origin: .....

First Language (language you speak at home): .....

How long have you been staying in the UK? .....months

How long have you been studying English? …….years

2) English Proficiency

a) Have you taken any English language tests? **YES/ NO** (please circle) If **YES**, please give details:

Test (e.g., FCE, IELTS, TOEFL)	**Date taken (DD/MM/YY, e.g., 01/12/05)**	**Grade or Score (if known)**		
		Overall	**Listening**	**Speaking**

3) Familiarity with English spoken by shared L1 speakers and non-shared L1 speakers (please circle)

a) I am familiar with English spoken by **shared L1 speakers** (please circle)

1. Strongly disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly agree

b) I am familiar with English spoken by **non-shared L1 speakers** (please circle)

1. Strongly disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly agree

Appendix 2

Listening Test (Pilot Study)

(Cambridge ESOL, 2009, pp. 22-27)

Part 1 (Questions 1 – 8)

Instructions: You will hear people talking in eight different situations. For questions 1-8, choose the best answer (A, B or C).

1) You overhear a young man talking about his first job. How did he feel in his first job?
   A. Bored B. Confused C. Enthusiastic

2) You hear a radio announcement about a dance company. What are listeners being invited to?
   A. A show B. A talk C. A party

3) You overhear a woman talking to a man about something that happened to her. Who was she?
   A. A pedestrian B. A driver C. A passenger

4) You hear a woman talking on the radio about her work making wildlife films. What is her main point?
   A. Being in the right place at the right time is a matter of luck.
   B. More time is spent planning than actually filming.
   C. It is worthwhile spending time preparing.

5) You hear part of a travel programme on the radio. Where is the speaker?
   A. Outside a cafe B. By the sea C. On a lake

6) You overhear a woman talking about a table-tennis in a sport shop. What does she want the shop assistant to do about her table-tennis table?
   A. Provide her with a new one.
   B. Have it put together for her.
   C. Give her the money back.

7) You hear part of an interview with a businessman. What is her business?
   A. Hiring out boats
   B. Hiring out caravans
   C. Building boats

8) You hear a man talking on the radio. Who is talking?

A. An actor B. A journalist C. A theatre-goer

Part 2 (Questions 9 – 18)

Instructions: You will hear a radio interview with Mike Reynolds, whose hobby is exploring underground places such as caves. For questions 9-18, complete the sentences.

Cavers explore undergrounf places such as mines and (9) ……..……

As well as caves. When cavers campl underground, they choose places which have (10) ……and …… available. In the UK, the place Mike likes best for caving is (11) ……..…… As a physical actvity, Mike compares caving to (12) …… Cavers can pay as much as £20 for a suitable (13) ……

Cavers can pay as much as £50 for the right kind of (14) ……..……., which is worn on the head. Mike recommends buying expensive (15)……..… to avoid having accidents. Caving is a sport for people of (16) ……and backgrounds. Some cavers in Britain are called "places of (17)……" The need for safety explains why people don't organise caving (18)……

Part 3 (Questions 19 – 23)

Instructions: You will hear five different people talking about their work on a cruise ship. For questions 19-23, choose from the list (A-F0 what each speaker says about their work. Use the letters only once. There is one extra letter which you do not need to use.

A One aaspet of my job is less interesting than others.

Speaker 1 (19) ☐

BMy job involves planning for the unexpected

Speaker 2 (20) ☐

CYou have to be socialable to do my job.

Speaker 3 (21) ☐

D I don't like routine in my working life.

Speaker 4 (22) ☐

E There's mot much work to do during the day.

Speaker 5 (23) ☐

F I provide passagers with a souvenir of their trip.

Part 4 (Questions 24 – 30)

Instructions: You will hear an interview with a man called Stan Leach who is talking about an adventure sports. For questions 24-30, choose the best answer (A, B or C).

24) Stan says the best thing about walking is that you can

A get fit by doing it.

B please yourself how you do it.

C do it on your own.

25) Stan's opinion on scambling is that

A people doing it may need to be accompanied.

B it is unsuitable for beginners.

C it is more exciting than walking.

26) What did Stan discover when he went climbing?

A It was not enjoyable.

B It was harder than he expected.

C It can be very frightening.

27) What does Stan say about mountain biking?

A Britain is not the best place for it.

B It is more expensive in Britain than elsewhere.

C It is best where there are lots of downhill slopes.

28) Stan's advice on scuba diving is that

A most of the courses for it are good.

B it is easier than it seems.

C you should think carefully before trying it.

29) What is Stan's view of skydiving?

A It is surprisingly popular.

B It is best when done in teams.

C Only certain types of people like it.

30) What does Stan say about canoeing?

A You can do it conditions that suit you.

B It is best at certain times of the year.

C There are few places in Britain to do it.

========== **THE END OF THE TEST** ==========

Appendix 3

Monologic Speaking Test

(Cambridge ESOL, 2009)

Instructions:

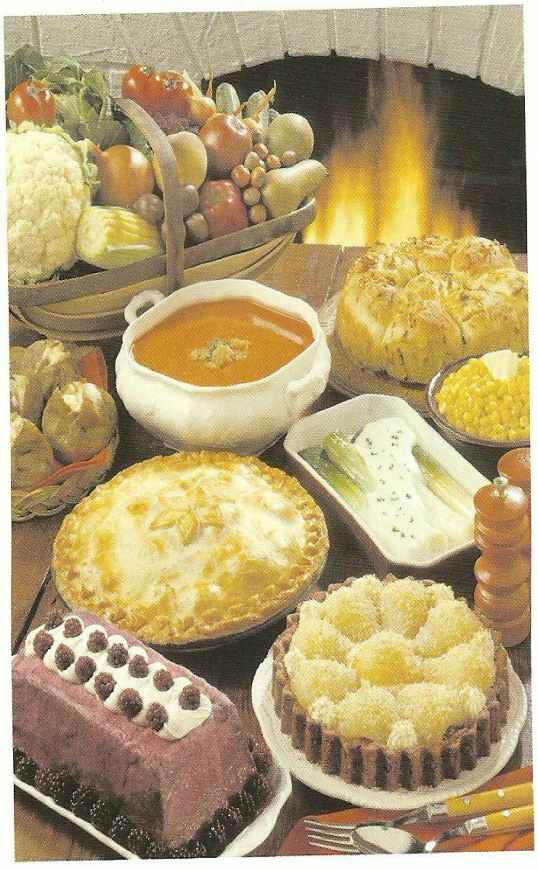
1 Please tell me about yourself, for example, your name, your hobby, for a minute.

2 Answers questions related to the given photos for 2 minutes.

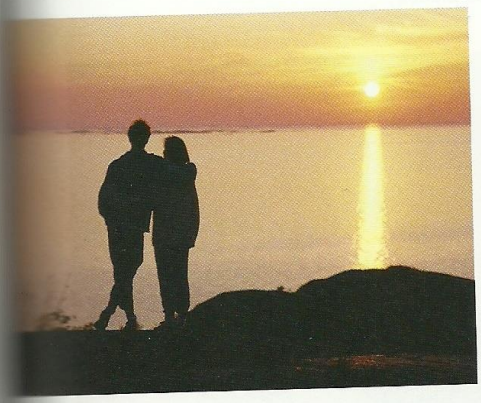
1) How important are these things for a happy life?
2) Which two are the most important?

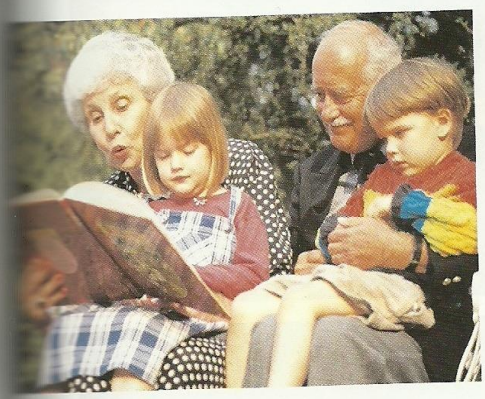
- How important are these things for a happy life?
- Which two are the most important?

4E



C14





C15

Appendix 4

Paired Speaking Tests

(Cambridge ESOL, 2009)

TASK A

Instructions:

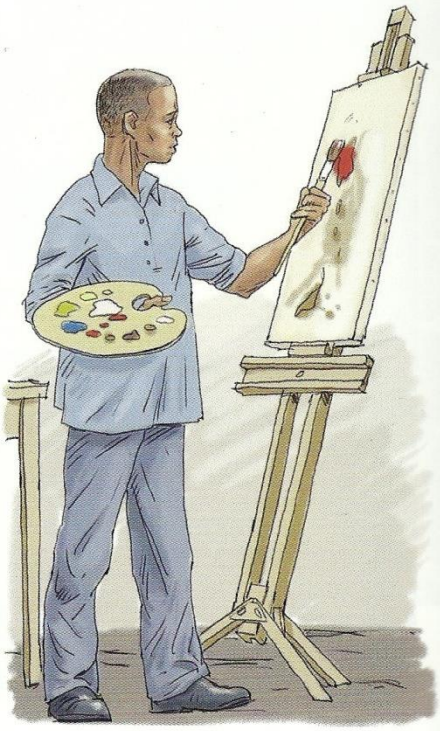
1 Please introduce yourselves to each other for one minute.

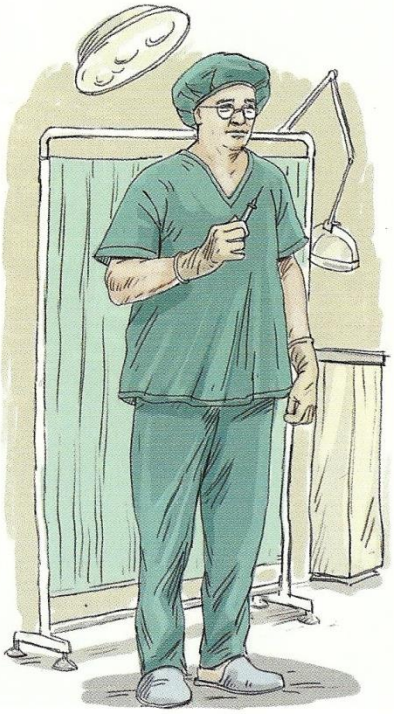
2 Answers questions by linking to the given photos. You have to discuss with your partner for 4 minutes.

1) How difficult is it to be successful in these professions?
2) In which profession is it most difficult to get to the top?

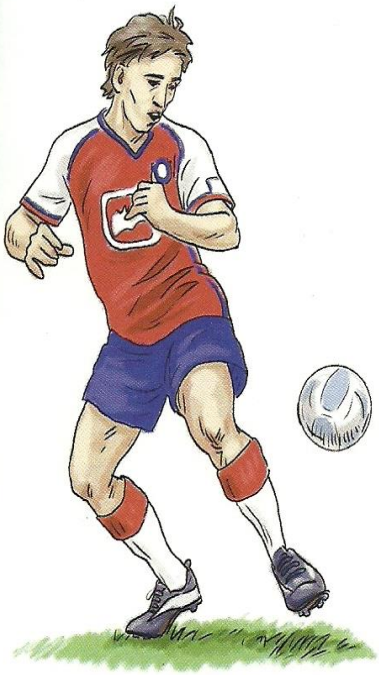
- How difficult is it to be successful in these professions?
- In which profession is it most difficult to get to the top?

3E





C10



C11

TASK B

Instructions:

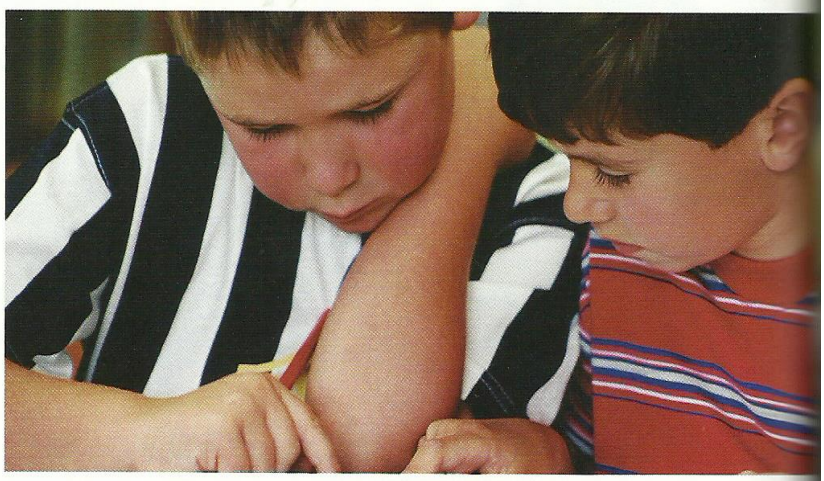
1 Please introduce yourselves to each other for one minute.

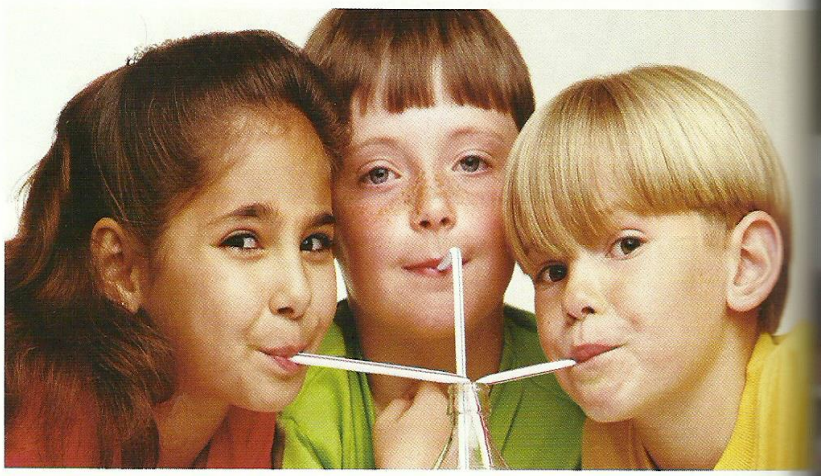
2 Answers questions by linking to the given photos. You have to discuss with your partner for 4 minutes.

1) What are the advantages of having friends?
2) In which situation are friends most important?

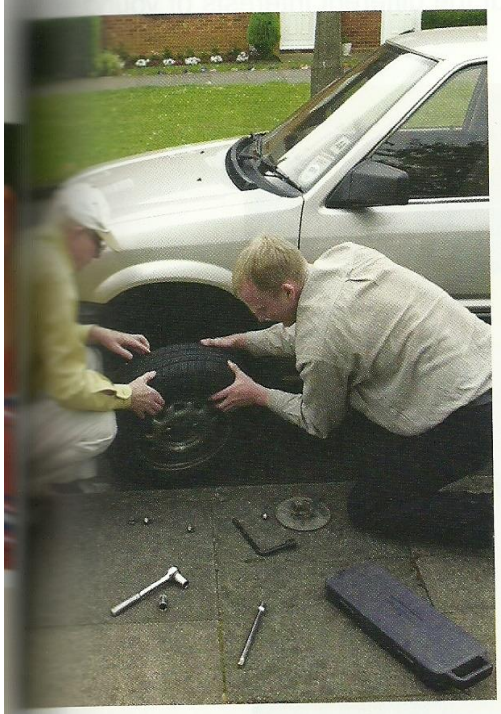
- What are the advantages of having friends?
- In which situation are friends most important?

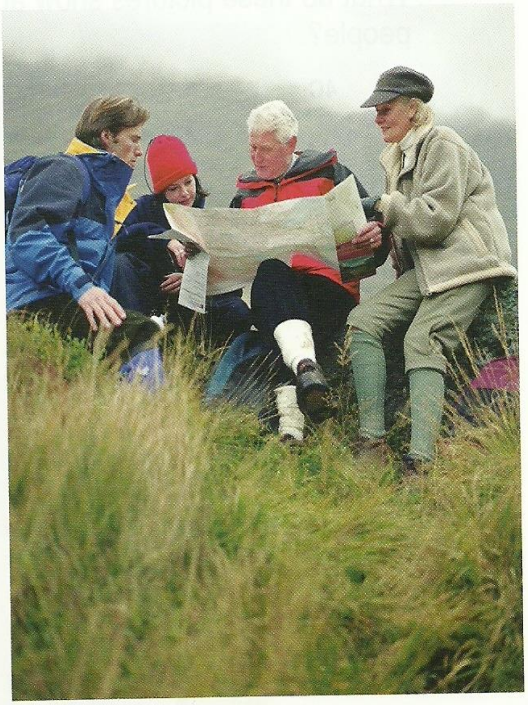
4E

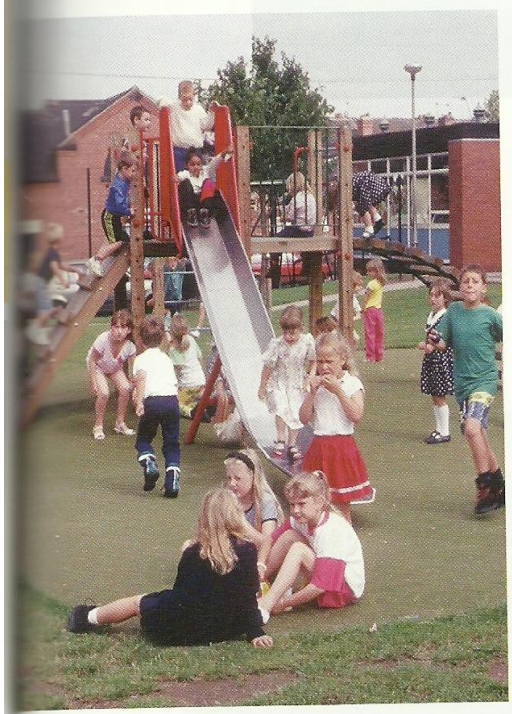


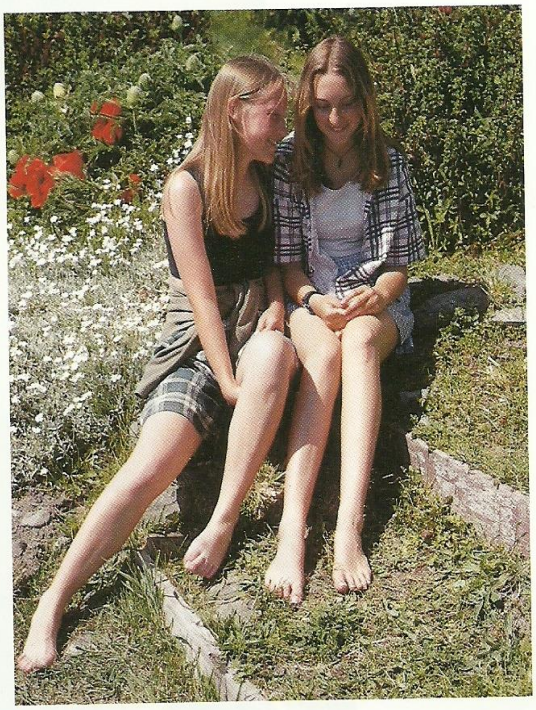


C14









Appendix 5

Speaking Test Criteria (Pilot Study)

(UCLES, 2009)

Interactive communication category is not included in the monologic speaking test. Therefore, there is no score for the interactive communication in the monologic speaking test.

Band	Grammar and vocabulary	Discourse management	Pronunciation	Interactive communication
5.0	Gammar is mostly accurate. Only minor errors occur. Uses appropriate and varied vocabulary in dealing with the tasks.	Uses wide range of linguistic resources to deal effectively with the tasks. Contributions are coherent and extended where appropriate.	Produces individual sounds well and speaks with appropriate intonation and stress. Although L1 accent may be evident, utterances are easily understood.	Demonstrates good interactive ability in carrying out the tasks. Is able to maintain effective communication with only natural hesitation while organising thoughts and shows sensitivity to turn-taking. Does not require assistance in acrrying out the tasks.
4.5				
4.0	Some features of 3 and some features of 5 in approximately equal measure.	Some features of 3 and some features of 5 in approximately equal measure.	Some features of 3 and some features of 5 in approximately equal measure.	Some features of 3 and some features of 5 in approximately equal measure.
3.5				
3.0	Grammar is sufficiently accurate. Uses appropriate vocabulary in dealing with the tasks.	Uses adequate range of linguistic resources to deal sufficiently well with the tasks. Contributions may occasionally be limited or lack coherence.	Produces individual sounds and prosodic features sufficiently well to be understood. L1 accent may cause occasional difficulty.	Has sufficient interactive ability to carry out the tasks. Maintains flow of language when carrying out the tasks although may occasioally lack sensitivity to turn taking and hesitation may occur while searching for language. Dose not require major assistance or prompting to carry out the tasks.
2.5				
2.0	Some features of 1 and some features of 3 in approximately equal measure.	Some features of 1 and some features of 3 in approximately equal measure.	Some features of 1 and some features of 3 in approximately equal measure.	Some features of 1 and some features of 3 in approximately equal measure.
1.5				
1.0	Grammar is mostly inaccurate. Major errors occur. Uses limited or inappropriate vocabulary in dealing with the tasks.	Range of linguistics resources is inadequate to deal with the tasks. Contributions are often minimal and lack coherence.	Produces some features of spoken English poorly that utterances are not easily understood. L1 accent puts sain on the listener.	Is only able to take part in the interaction for some of the time. Cannot maintain flows of language and hestations demand patience of the listener. Requires major prompting and assistance. Produces inappropriate or irrelevant responses.

Appendix 6

Transcription Symbols

(Atkinson and Heritage, 1984)

(0.5)	Number of a time gap in tenths of a second
(.)	A pause which is less than two-tenths of a second
=	A latch between utterances
[	An onset of overlapping talk
.hh	A speaker's in-breath
hh	A speaker's out-breath
(( ))	A non-verbal activity
-	A sharp cut-off of the prior word or sound
:	Stretch sound or letter
( )	An unclear fragment on the tape
(guess)	The transcriber's best guess at an unclear utterance
.	A stopping fall in tone
,	A continuing intonation
?	A rising intonation
Underlining	An emphasised word or sound
◦ ◦	The talk quieter than the surrounding
> <	The talk quicker than the surrounding
< >	The talk more slow down than the surrounding
Hah, huh, heh	Laughing

Appendix 7

Speaking Test Criteria (Main Study)

(UCLES, 2012)

Interactive communication category is no included in the monologic speaking test. Therefore, there is no score for the interactive communication in the monologic speaking test.

Band	Grammar and vocabulary	Discourse management	Pronunciation	Interactive communication
5	Shows a good degree of control of a range of simple and some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a wide range of familiar topics.	Produces extended stretches of language with very little hesitation. Contributions are relevant and there is a clear organisation of ideas. Uses a range of cohesive devices and discourse markers.	Is intelligible. Intonation is appropriate. Sentence and word stress is accurately placed. Individual sounds are articulated clearly.	Initiates and responds appropriately, linking contributions to those of other speakers. Maintains and develops the interaction and negotiates towards an outcome.
4	*Performance shares features of Bands 3 and 5.*			
3	Shows a good degree of control of simple gramatical forms, and attempts some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a range of familiar topics.	Produces extended stretches of language despite some hesitation. Contributions are relevant and there is very little repetition. Uses a range of cohesive devices.	Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly.	Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.
2	*Performance shares features of Bands 1 and 3.*			
1	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary when talking about everyday situations.	Produces responses which are extended beyond short phrases, despite hesitation. Contributions are mostly relevant, depite some repetition. Uses basic cohesive devices.	Is mostly intelligible , and has some control of phonological features at both utterance and word levels.	Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.
0	*Performance below Band 1.*			

Appendix 8

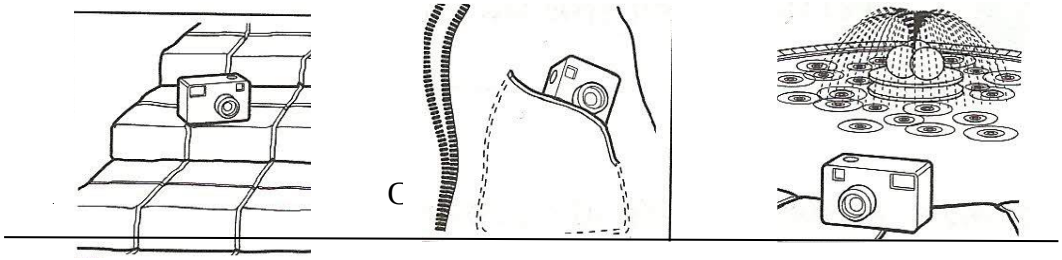
Listening Test (Main Study)

(Cambridge ESOL, 2008, 2009)

Part 1 (Questions 1-5)

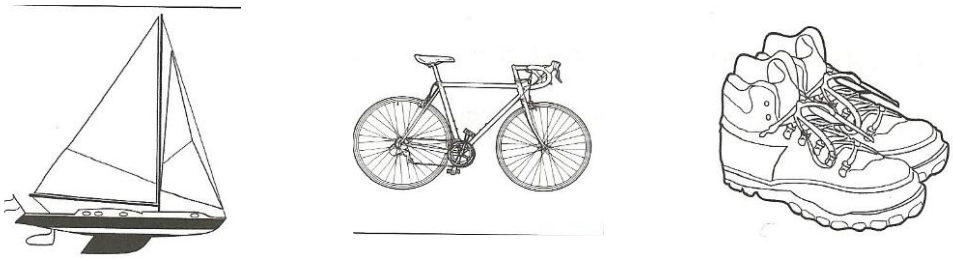
Instructions: There are seven questions in this part. For each question there are three pictures and a short recording. Choose the correct picture from the given choices (A, B or C).

Example: Where did the man leave this camera?



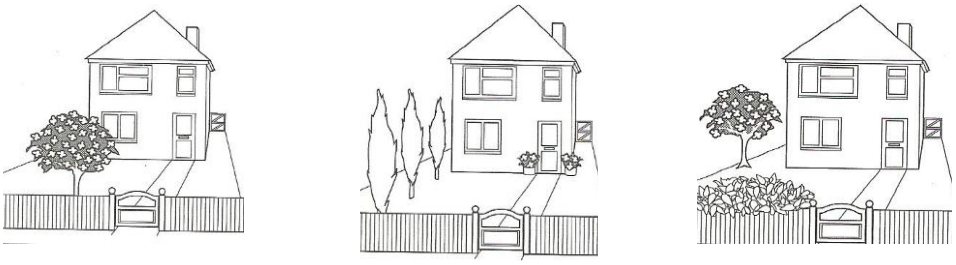
A **A** B C

1) Which activity will the family do this year?



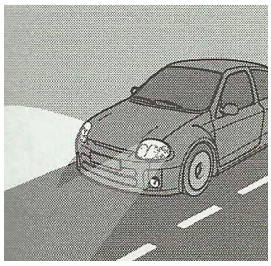
A B C

2) Which is the woman's house?

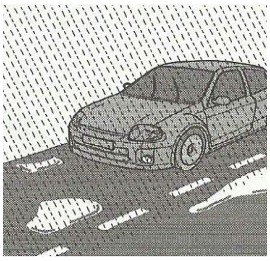


A B C

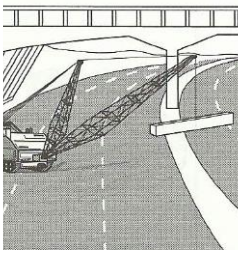
3) Why did drivers have problems this morning?



A

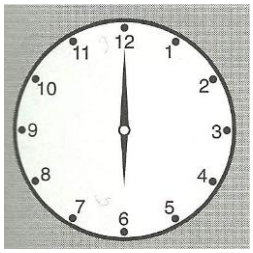


B

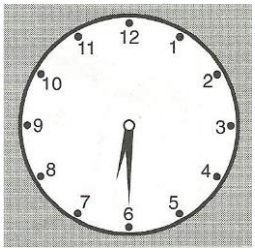


C

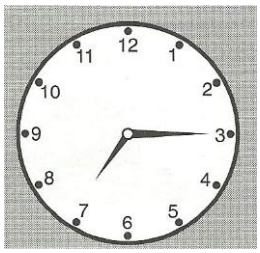
4) What time will Robin leave the house?



A

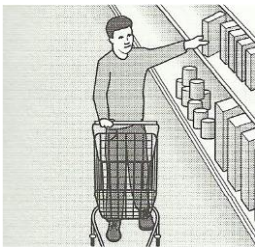


B



C

5) What did Simon do this morning?



A

B

C

Part 2 (Questions 6 - 11)

Instructions: You will hear an interview with Angela Morgan, who has recently flown around the world in a helicopter. For each question, choose the correct answer (A, B or C).

6) The main reason for Angela's trip was to

A. make money for her business

B. make money for other people

C. have an exciting adventure.

7) What does Angela say about her life now?

A. She feels much older.

B. She likes to be active and busy.

C. She is lonely without her children.

8) When Angela had flying lessons

A. her course lasted five months.

B. her husband took lessons as well.

C. she got to know her teacher well.

9) During the trip, Angela and her teacher

A. did very little sightseeing.

B. carried all the water they needed.

C. had engine problems several times.

10) What did Angela enjoy most about the trip?

A. flying at night

B. walking in the desert

C. watching the changes in the scenery

11) What did Angela miss most while she was away?

A. modern bathrooms

B. regular exercise

C. interesting entertainment

Part 3 (Questions 12 – 17)

Instructions: Look at the six sentences for this part. You will hear a conversation between a boy, Tom, and his sister, Clare, about school. Decide if each sentence is correct or incorrect. If it is correct, put a tick (√) in the box under **A** for **YES**. If it is not correct, put a tick (√) in the box under **B** for **NO**.

	A	**B**
	YES	**NO**
12) Clare thinks their father will be pleased by Tom's news.	☐	☐
13) Tom believes he can manage both swimming and school work.	☐	☐
14) Tom's teacher thinks Tom is clever.	☐	☐
15) Tom dislikes doing maths.	☐	☐
16) Clare thinks it is a bad idea to take a friend's advice.	☐	☐
17) Tom finally realizes he will need his father's agreement to his plans.	☐	☐

Part 4 (Questions 18 - 25)

Instructions: You will hear people talking in eight different situations. For questions 18 - 25 choose the best answer (A, B or C).

18) You overhear a young man talking about his first job. How did he feel in his first job?

A. Bored B. Confused C. Enthusiastic

19) You hear a radio announcement about a dance company. What are listeners being invited to?

A. A show B. A talk C. A party

20) You overhear a woman talking to a man about something that happened to her. Who was she?

A. A pedestrian B. A driver C. A passenger

21) You hear a woman talking on the radio about her work making wildlife films. What is her main point?

A. Being in the right place at the right time is a matter of luck.

B. More time is spent planning than actually filming.

C. It is worthwhile spending time preparing.

22) You hear part of a travel programme on the radio. Where is the speaker?

A. Outside a cafe B. By the sea C. On a lake

23) You overhear a woman talking about a table-tennis table in a sport shop. What does she want the shop assistant to do about her table-tennis table?

A. Provide her with a new one.

B. Have it put together for her.

C. Give her the money back.

24) You hear part of an interview with a businessman. What is her business?

D. Hiring out boats

E. Hiring out caravans

F. Building boats

25) You hear a man talking on the radio. Who is talking?

A. An actor B. A journalist C. A theatre-goer

Part 5 (Questions 26 – 30)

Instructions: You will hear five different people talking about their work on a cruise ship. For questions 26-30, choose from the list (A-F0 what each speaker says about their work. Use the letters only once. There is one extra letter which you do not need to use.

A One aaspet of my job is less interesting than others.

Speaker 1 (26) ☐

B My job involves planning for the unexpected

Speaker 2 (27) ☐

C You have to be socialable to do my job.

Speaker 3 (28) ☐

D I don't like routine in my working life.

Speaker 4 (29) ☐

E There's mot much work to do during the day.

Speaker 5 (30) ☐

FI provide passagers with a souvenir of their trip.

Part 6 (Questions 31 – 37)

Instructions: You will hear an interview with a man called Stan Leach who is talking about adventure sports. Choose the best answer (A, B or C).

31) Stan says that the best thing about walking is that you can

A. get fit by doing it.

B. please yourself how you do it.

C. do it on your own.

32) Stan's opinion on scrambling is that

A. people doing it may need to be accompanied.

B. it is unsuitable for beginners.

C. it is more exciting than walking.

33) What did Stan discover when he went climbing ?

A. It was not enjoyable.

B. It was harder than he expected.

C. It can be very frightening.

34) What does Stan say about mountain biking?

A. Britain is not the best place for it.

B. It is more expensive in Britain than elsewhere.

C. It is best where there are lots of downhill slopes.

35) Stan's advice on scuba diving is that

A. most of the courses for it are good.

B. it is easier than it seems.

C. you should think carefully before trying it

36) What is Stan's view of skydiving?

A. It is surprisingly popular.

B. It is best when done in teams.

C. Only certain types of people like it.

37) What does Stan say about canoeing?

A. You can do it in conditions that suit you.

B. It is best at certain times of the year.

C. There are few places in Britain to do it.

========== **THE END OF THE TEST** ==========

Appendix 9

Reliability of the Listening Test

39 test- items (Try out before the main data collection)

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.905	.900	39

Item Statistics			
	Mean	Std. Deviation	N
a01p01	.9333	.25371	30
a02p02	.6333	.49013	30
a03p03	.8667	.34575	30
a04p04	.7667	.43018	30
a05p05	.4333	.50401	30
a06p06	.8333	.37905	30
a07p07	.8333	.37905	30
a08p08	.6333	.49013	30
a09p09	.4333	.50401	30
a10p10	.6333	.49013	30
a11p11	.5333	.50742	30
a12p12	.7667	.43018	30
a13p13	.7333	.44978	30
a14p20	.7667	.43018	30
a15p21	.6333	.49013	30
a16p22	.7333	.44978	30
a17p23	.7667	.43018	30
a18p24	.7667	.43018	30
a19p25	.7333	.44978	30
a20f01	.6000	.49827	30
a21f02	.5333	.50742	30
a22f03	.4667	.50742	30
a23f04	.3333	.47946	30
a24f05	.4333	.50401	30
a25f06	.3667	.49013	30
a26f07	.3333	.47946	30
a27f08	.4333	.50401	30
a28f19	.5333	.50742	30
a29f20	.5333	.50742	30

a30f21	.4000	.49827	30
a31f22	.4000	.49827	30
a32f23	.3333	.47946	30
a33f24	.3667	.49013	30
a34f25	.3667	.49013	30
a35f26	.6333	.49013	30
a36f27	.3667	.49013	30
a37f28	.4667	.50742	30
a38f29	.3333	.47946	30
a39f30	.6333	.49013	30

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
a01p01	21.3667	73.068	-.052	.	.907
a02p02	21.6667	69.402	.400	.	.903
a03p03	21.4333	73.289	-.085	.	.908
a04p04	21.5333	69.499	.449	.	.902
a05p05	21.8667	70.120	.300	.	.904
a06p06	21.4667	69.982	.439	.	.902
a07p07	21.4667	71.499	.197	.	.905
a08p08	21.6667	69.678	.365	.	.903
a09p09	21.8667	70.464	.259	.	.905
a10p10	21.6667	66.575	.762	.	.898
a11p11	21.7667	65.978	.809	.	.897
a12p12	21.5333	69.982	.381	.	.903
a13p13	21.5667	69.082	.485	.	.902
a14p20	21.5333	70.878	.255	.	.905
a15p21	21.6667	69.540	.382	.	.903
a16p22	21.5667	70.047	.353	.	.903
a17p23	21.5333	70.326	.332	.	.904
a18p24	21.5333	69.499	.449	.	.902
a19p25	21.5667	69.702	.400	.	.903
a20f01	21.7000	69.321	.402	.	.903
a21f02	21.7667	67.771	.584	.	.900
a22f03	21.8333	68.006	.555	.	.901
a23f04	21.9667	70.309	.294	.	.904
a24f05	21.8667	69.154	.417	.	.903
a25f06	21.9333	70.478	.266	.	.905

a26f07	21.9667	68.447	.533	.	.901
a27f08	21.8667	70.326	.275	.	.905
a28f19	21.7667	66.599	.731	.	.898
a29f20	21.7667	69.633	.356	.	.904
a30f21	21.9000	69.541	.375	.	.903
a31f22	21.9000	68.645	.486	.	.902
a32f23	21.9667	70.102	.321	.	.904
a33f24	21.9333	68.340	.534	.	.901
a34f25	21.9333	67.444	.649	.	.899
a35f26	21.6667	68.506	.513	.	.901
a36f27	21.9333	68.823	.473	.	.902
a37f28	21.8333	68.626	.479	.	.902
a38f29	21.9667	70.171	.312	.	.904
a39f30	21.6667	67.333	.663	.	.899

Thirty-seven items (deleted items 1 and 3)

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.910	.909	37

Item Statistics			
	Mean	Std. Deviation	N
a02p02	.6333	.49013	30
a04p04	.7667	.43018	30
a05p05	.4333	.50401	30
a06p06	.8333	.37905	30
a07p07	.8333	.37905	30
a08p08	.6333	.49013	30
a09p09	.4333	.50401	30
a10p10	.6333	.49013	30
a11p11	.5333	.50742	30
a12p12	.7667	.43018	30
a13p13	.7333	.44978	30
a14p20	.7667	.43018	30
a15p21	.6333	.49013	30
a16p22	.7333	.44978	30
a17p23	.7667	.43018	30
a18p24	.7667	.43018	30
a19p25	.7333	.44978	30
a20f01	.6000	.49827	30

a21f02	.5333	.50742	30
a22f03	.4667	.50742	30
a23f04	.3333	.47946	30
a24f05	.4333	.50401	30
a25f06	.3667	.49013	30
a26f07	.3333	.47946	30
a27f08	.4333	.50401	30
a28f19	.5333	.50742	30
a29f20	.5333	.50742	30
a30f21	.4000	.49827	30
a31f22	.4000	.49827	30
a32f23	.3333	.47946	30
a33f24	.3667	.49013	30
a34f25	.3667	.49013	30
a35f26	.6333	.49013	30
a36f27	.3667	.49013	30
a37f28	.4667	.50742	30
a38f29	.3333	.47946	30
a39f30	.6333	.49013	30

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
a02p02	19.8667	69.913	.400	.	.908
a04p04	19.7333	69.926	.461	.	.907
a05p05	20.0667	70.823	.277	.	.910
a06p06	19.6667	70.506	.437	.	.907
a07p07	19.6667	72.023	.197	.	.910
a08p08	19.8667	70.189	.366	.	.908
a09p09	20.0667	70.823	.277	.	.910
a10p10	19.8667	67.085	.760	.	.903
a11p11	19.9667	66.516	.804	.	.902
a12p12	19.7333	70.547	.373	.	.908
a13p13	19.7667	69.495	.498	.	.907
a14p20	19.7333	71.444	.248	.	.910
a15p21	19.8667	70.189	.366	.	.908
a16p22	19.7667	70.461	.367	.	.908
a17p23	19.7333	70.823	.335	.	.909
a18p24	19.7333	69.995	.452	.	.907

a19p25	19.7667	70.116	.413	.	.908
a20f01	19.9000	69.817	.404	.	.908
a21f02	19.9667	68.240	.588	.	.905
a22f03	20.0333	68.516	.554	.	.906
a23f04	20.1667	70.833	.293	.	.909
a24f05	20.0667	69.720	.411	.	.908
a25f06	20.1333	71.085	.255	.	.910
a26f07	20.1667	69.040	.522	.	.906
a27f08	20.0667	70.754	.286	.	.909
a28f19	19.9667	67.068	.735	.	.903
a29f20	19.9667	70.033	.370	.	.908
a30f21	20.1000	70.024	.379	.	.908
a31f22	20.1000	69.128	.489	.	.907
a32f23	20.1667	70.557	.328	.	.909
a33f24	20.1333	68.878	.530	.	.906
a34f25	20.1333	67.982	.645	.	.904
a35f26	19.8667	68.947	.521	.	.906
a36f27	20.1333	69.361	.469	.	.907
a37f28	20.0333	69.137	.478	.	.907
a38f29	20.1667	70.695	.311	.	.909
a39f30	19.8667	67.844	.662	.	.904

Appendix 10

Questionnaire Results for the Main Study

11.1 Test-takers' demographic information

The test-takers' age, length of stay in the UK and number of years of studying English are summarised in this section. There were 20 Urdu L1 and 20 Thai L1 test-takers. Both Urdu and Thai L1 test-taker groups consisted of 10 males and 10 females. Descriptive statistics for the 40 test-takers are illustrated first, followed by those for the 20 Urdu and 20 Thai L1 test-takers.

Table 11.1: Statistics for test-takers' age, length of stay in the UK and length of studying English (N=40) – main study

	Min	Max	Mean	Median	SD
Age (years)	23	34	27.20	27.00	2.84
Length of stay in the UK (months)	1	18	6.60	4.50	4.35
Length of studying English (years)	5	22	13.53	12.00	4.38

11.1.1 Test-takers' demographic information divided by their L1s (Urdu and Thai)

Next, the test-takers' demographic information is compared between 20 Urdu L1 (10 males and 10 females) test-takers and 20 Thai L1 (10 males and 10 females) test-takers.

Table 11.2: Statistics for Urdu (N=20) and Thai (N=20) L1 test-takers' age, length of stay in the UK and length of studying English – main study

	L1	Min	Max	Mean	Median	SD
Age (years)	Urdu	23	33	27.40	28.00	2.89
	Thai	23	34	27.00	27.00	2.85
Length of stay in the UK (months)	Urdu	1	12	5.05	4.00	3.28
	Thai	1	18	8.15	8.50	4.79
Length of studying English (years)	Urdu	5	22	12.35	11.50	4.85
	Thai	12	22	14.70	12.00	3.60

As summarised in Table 11.2, the minimum age of the Urdu and Thai L1 test-takers was the same, 23 years old, while the maximum age of the Thai L1 test-takers (Mean=27.00, SD=2.85) was older than that of the Urdu L1 test-takers (Mean=27.40, SD=2.89). However, they were in a very similar age range. The

length of time the Urdu L1 test-takers had been in the UK ranged from one month to one year (Mean=5.05, SD=3.28). The length of time the Thai L1 test-takers had been in the UK ranged from one month to a year and a half (Mean=8.15, SD=4.79). Thr average length of time that the Thai L1 test-takers had spent in the UK was longer than for Urdu L1 test-takers, for whom it was approximately 3 months. The minimum length of time the Thai L1 test-takers had been studying English was longer than for the Urdu L1 test-takers. The average length of time that the Thai L1 test-takers had been studying English was longer than for the Urdu L1 test-takers approximately 2 years.

11.2 Test-takers' English proficiency based on the IELTS scores

A measure of all the test-takers' (20 Urdu and 20 Thai L1) English proficiency based on the IELTS examination scores was provided for these test-takers. The statistics for overall, speaking and listening proficiency based on the IELTS scores are illustrated in Table 11.3.

Table 11.3: Statistics for test-takers' overall, speaking and listening proficiency based on the IELTS scores (N=40) – main study

	Min	Max	Mean	Median	SD
IELTS overall	4.5	5.5	5.33	5.50	.27
IELTS speaking	5.0	6.0	5.61	5.50	.35
IELTS listening	5.0	6.5	5.28	5.00	.39

As illustrated in Table 11.3, 40 test-takers' (both Urdu and Thai L1) English proficiency based on the IELTS overall scores ranged from Band 4.5 to Band 5.5 (Mean=5.33, SD=.27). Their IELTS speaking scores ranged from Band 5.0 to Band 6.0 (Mean=5.61, SD=.35) and their listening proficiency ranged from Band 5.0 to Band 6.5 (Mean=5.28, SD=.39). Judging from the range and mean scores, the participants seemed to have a slightly lower score in listening than speaking.

Statistical information for Urdu and Thai L1 test-takers' English proficiency, speaking proficiency and listening proficiency based on IELTS examination scores are provided and compared in the following section.

11.2.1 Urdu (N=20) and Thai L1 (N=20) test-takers' overall, speaking and listening proficiency based on the IELTS examination scores

As shown in Table 11.4, Urdu L1 test-takers' IELTS overall scores ranged from Band 5.0 to Band 5.5 (Mean=5.40, SD=.21), while Thai L1 test-takers' overall scores ranged from Band 4.5 to Band 5.5 (Mean=5.25, SD=.30). Urdu L1 test-takers' IELTS speaking scores ranged from Band 5.0 to Band 6.0 (Mean=5.70, SD=.30), and their listening scores ranged from Band 5.0 to 6.5 (Mean=5.40, SD=.48). Thai L1 test-takers' speaking scores ranged from Band 5.0 to 6.0 (Mean=5.53, SD=.38), and their listening scores from Band 5.0 to 5.5 (Mean=5.15, SD=.24).

Table 11.4: Statistics for Urdu (N=20) and Thai L1 (N=20) test-takers' overall, speaking and listening proficiency based on the IELTS scores – Main study

	L1	Min	Max	Mean	Median	SD
IELTS overall	Urdu	5.0	5.5	5.40	5.50	.21
	Thai	4.5	5.5	5.25	5.50	.30
IELTS speaking	Urdu	5.0	6.0	5.70	5.50	.30
	Thai	5.0	6.0	5.53	5.50	.38
IELTS listening	Urdu	5.0	6.5	5.40	5.25	.48
	Thai	5.0	5.5	5.15	5.00	.24

As shown through the average scores, both groups of test-takers had similar overall, speaking and listening proficiency levels, although Urdu L1 test-takers had very slightly higher means across all scores.

11.3 Test-takers' familiarity with the English spoken by shared and non-shared L1 speakers

To rate the test-takers' opinions on their familiarity with the English spoken by shared and non-shared L1 speakers, Likert-scale questions were employed. The test-takers were asked to rate their familiarity ranging from 1 to 5 (1=strongly disagree, 2=disagree, 3=neutral opinion, 4=agree and 5=strongly agree). This section presents the results for the 20 Urdu and Thai L1 test-takers' opinions on their familiarity with the English spoken by shared L1 speakers and by non-shared L1 speakers (i.e., either Urdu or Thai L1 speakers).

Table 11.5: Statistics for Urdu (N=20) and Thai L1 (N=20) test-takers' opinions on their familiarity with the English spoken by Urdu and Thai L1 speakers

L1	Opinion	Min	Max	Mean	Median	SD
Urdu	Familiar with the English spoken by **Urdu** L1 speakers	3	5	4.20	4.00	.62
	Familiar with the English spoken by **Thai** L1 speakers	2	5	3.60	4.00	.75
Thai	Familiar with the English spoken by **Urdu** L1 speakers	1	4	2.25	2.00	.79
	Familiar with the English spoken by **Thai** L1 speakers	1	5	3.85	4.00	1.04

Urdu L1 test-takers' opinion on their familiarity with the English spoken by Urdu L1 speakers ranged from 3 (neutral) to 5 (strongly agree), with a mean score of 4.20 (SD=.62). Their opinion on their familiarity with the English spoken by Thai L1 speakers ranged from 2 (disagree) to 5 (strongly agree) and the mean score was 3.60 (SD=.75),

The ratings for the Thai L1 test-takers' familiarity with the English spoken by Urdu L1 speakers ranged from 1 (strongly disagree) to 4 (agree), and the mean score was 2.25 (SD=.79). Their ratings for the English spoken by Thai L1 speakers varied from 1 (strongly disagree) to 5 (strongly agree), with a mean of 3.85 (SD=1.04).

Frequency information for Urdu and Thai L1 test-takers' opinions on their familiarity with the English spoken by Thai L1 speakers is shown in Table 11.6.

Table 11.6: Frequency information for Urdu (N=20) and Thai L1 (N=20) test-takers' opinions on their familiarity with the English spoken by Urdu and Thai L1 speakers – main study

		Urdu L1 test-takers' opinion on their familiarity with the English spoken by				**Thai L1** test-takers' opinion on their familiarity with the English spoken by			
		Frequency		**Percentage**		**Frequency**		**Percentage**	
		Urdu	**Thai**	**Urdu**	**Thai**	**Urdu**	**Thai**	**Urdu**	**Thai**
Valid	Strongly disagree	-	-	-	-	3	1	15	5
	Disagree	-	2	-	10	10	-	50	-
	Neutral	2	5	10	25	6	6	30	30
	Agree	12	12	60	60	1	7	5	35
	Strongly agree	6	1	30	5	-	6	-	30
	Total	20	20	100	100	20	20	100	100

Figures 11.1 and 11.2 present boxplots for Urdu and Thai L1 test-takers' opinion on their familiarity with English spoken by Urdu and Thai L1 speakers.

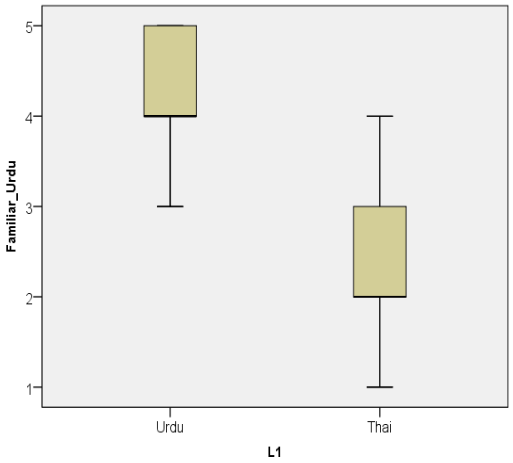


Figure 11.1: Boxplot for Urdu (N=20) and Thai L1 (N=20) test-takers' opinions on their familiarity with the English spoken by Urdu L1 speakers

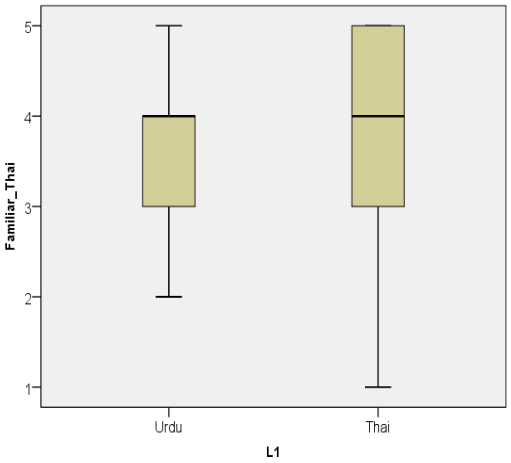


Figure 11.2: Boxplot for Urdu (N=20) and Thai L1 (N=20) test-takers' opinions on their familiarity with the English spoken by Thai L1 speakers

To investigate the differences between Urdu and Thai L1 test-takers' opinions on their familiarity with the English spoken by shared and non-shared L1 speakers (i.e., Urdu and Thai L1 speakers), two non- parametric Wilcoxon signed-rank tests were used. The results are shown in Table 11.7.

Table 11.7: Differences between Urdu and Thai L1 test-takers' opinions on their familiarity with the English spoken by Urdu and Thai L1 speakers

Source of information	Speakers	Mean	Median	SD	Wilcoxon
Urdu L1 test-takers' opinion on their familairity with the English spoken by:	Urdu L1	4.20	4.00	.62	Z=-2.65 p= .01
	Thai L1	3.60	4.00	.75	
Thai L1 test-takers' opinion on their familiarity with the English spoken by:	Urdu L1	2.25	2.00	.79	Z=-3.67 p=.00
	Thai L1	3.85	4.00	1.04	

Urdu L1 test-takers reported that they were significantly more familiar with the English spoken by Urdu L1 speakers (Mean=4.20) than Thai L1 test-takers (Mean=2.25). Similarly, Thai L1 test-takers indicated that they were significantly more familiar with the English spoken by Thai L1 speakers (Mean=3.85) than Urdu L1 test-takers (Mean=3.60), although their ratings were in general lower than the Urdu L1 speakers' ratings across both the categories.

Appendix 11

Coding Scheme

Test-takers

1 Communication patterns in shared and non-shared L1 pairs

1.1 Supplying relevant vocabulary

A test-taker provides a relevant word or phrase while listening to a partner.

Example:

A: because ah:: it's like that (.) when you er::: when you it's
like (0.3) when you (.) find the burglar
B: ah::: yes ((nodding head))
A: burglar yeah so (0.3) he come to er::: (.) saa- ah:: (1.0)
→**B:** steal
A: steal some your phone or wallet
B: mm::

1.2 Demonstrating comprehension

A test-taker responds to a partner's message with a relevant contribution.

Example:

A: and they also can help you in a (.) difficult situation (0.9)
[some problems, (1.2) secret one ((raising hand)) [huh huh huh
→ **B:** [yeah [yeah yeah
→ secret one that's why I'm saying before that .hh (0.3) er: the friends
→ that (0.3) er they help us in er (0.5) ((moving hands)) the thing
→ that we can't say to our parents [or some other guys, .hh you can (.) =
B: [yeah

1.3 Back-channelling

A listener test-taker produces a speech sound in order to provide supporting feedback to a speaker (Ducasse, 2010; Ducasse and Brown, 2009), for example, *uh huh*, *yeah*, *yes*, *mm*.

Example:

A: … and also they .tch! (0.7) they can (.) help (0.3)
help to study to get the good grades [or or even give (0.3) give a
→ **B:** [yeah yeah yeah
A: hand (h) set (huh huh huh) of the answer ha ha ha

1.4 Attempting to understand a partner completely

When a communication problem occurs, a test-taker tries to understand a partner completely by using various strategies, e.g., confirmation check (e.g., *right?*, *you*

mean.., you think…) and clarification request (e.g., *What do you mean?*, *What?*, *Pardon?*, *Sorry?*).

Example:

A: so that the most difficult is this (.) this one.
B: yeah more responsibility with (.) the patients [too.
A: [yeah only one percent (0.7) can (0.9) er:: handle with the patient (2.5) which is ill.
→ **B:** so you think that mm:: ah surgeon (1.3) perhaps mm:: is the most difficult (.) profession to- to to get to the top, right?
A: yeah it's most difficult (0.7) according to my opinion.
B: ok ((nodding head))

1.5 Understanding an unclear utterance and incorrect word use

A test-taker can understand their partner even he/she produces an unclear utterance, incorrect words and wrong grammar.

Example:

→ **A:** how did your close- close friends?
B: I study together about the (0.3) more than more than ten years

1.6 Misunderstanding because of cultural background difference

A test-taker misunderstands his/her partner because they do not share the same cultural background.

Example:

A: and er: (.) ha ha i have a .hh vey short list of friends (.) yeah i only have two or three friends ha ha [and the mm:: (.) from my [part
B: [uh huh [you- you can do the party a lot man [yeah if you want to make a lot of friends [ha ha ha ha
→ **A:** [ha ha ha [yeah ha ha ha
→ actually problem's that i'm not good at party [ha ha ha
B: [oh yes
A: er:: i have not having some (0.3) lots of the friends ((moving hands)) [and (.)
B: [mm::
A: but er:: (.) in my start (0.3) [like ah:: i am here as a new
B: [uh huh
A: here [er:: one month
B: [uh huh ago in the uk er::: and er:: i also (0.3) er: get some (.) .hh kind of (.) the mm::: (.) lack of confident [also having some like (0.3) to er::
A: [uh huh uh huh ((nodding head))
B: is having problem ((moving hands)) to make with er: (.) the friends

[ev:ery friend and
A: [uh huh
B: what's happening
A: right uh huh
B: and er::: [.hh i must
A: [so let's go to the next hih hih hih

Note: A is Buddhist. B is Muslim.

A suggests B to arrange a party. B imagines that it must have alcohol in a party which is prohibited according to his religious beliefs. Therefore, B tries to explain his reasons only very implicitly that he cannot do it because he does not have skill in arranging parties and does not have self-confidence

2 Communication problems

2.1 Non-engagement

A test-taker does not participate in a conversation. He/she is quiet, answers a Wh-question with back-channelling or suddenly shifts to a new topic without extending a partner's idea.

Example:

A: what- >what do you think yourself which professions< is the most difficult to achieve (.) and .hh ah: (0.3) >what do you think is the most difficult< that level of level of high? .hh so what do you think of different professions here? ((pointing at pictures)) (.) so after that >I'll let you know about my my opinion< for- for for in my- my opinion what what professions to be difficult for me. (0.3) so (.) ((looking at a partner)) what what do you think?
(0.5)
→ **B:** yeah
A: ah: if you look at the first picture, what wha- ah what (.) profession is more difficult? ((looking at a partner))

2.2 Miscommunication

A test-taker provides an irrelevant response, partly relevant response, clarification request (e.g., *What do you mean?, What?, Pardon?, Sorry?*) or confirmation check.

Example:

A: er:: what is your profession?
→ **B:** .hh for me hh i'm not good at anything